Content Facets for Individual Information Needs in Media

Dissertation submitted to the
Graz University of Technology,
Faculty of Computer Science,
for the attainment of the degree of
Doctor of Engineering Sciences (Dr. techn.)

by

Elisabeth Lex
October 5, 2011

Thesis supervisors
First Advisor: Prof. Dr. Stephanie Lindstaedt
Second Advisor: Prof. Dr. Harald Kosch
Advisor: Dr. Michael Granitzer

Deutsche Fassung:
Beschluss der Curricula-Kommission für Bachelor-, Master- und Diplomstudien vom 10.11.2008
Genehmigung des Senates am 1.12.2008

EIDESSTATTLICHE ERKLÄRUNG

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommene Stellen als solche kenntlich gemacht habe.

Graz, am 05.09.2011 .....

.....
(Unterschrift)

Englische Fassung:

STATUTORY DECLARATION

I declare that I have authored this thesis independently, that I have not used other than the declared sources / resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

05.09.2011
.....
date

.....
(signature)

Abstract

The amount of content published in traditional media is huge and steadily growing. Additionally, social media gained momentum since people use blogs to share comments and opinions to current events. However, blog content is questionable in respect to quality since blogs are neither reviewed nor edited. From the media consumer perspective, navigating the haystack of information produced by media as well as finding content that meets ones quality demands is challenging. This challenge is the key motivation for this thesis: to support media consumers to filter media content by content facets capturing topical information needs as well as content quality aspects.

For this, two types of content facets are suggested: (i) topic oriented and (ii) topic independent quality related content facets. First, for each facet type (i) and (ii), concrete content facets are proposed. These content facets are then formally defined. Second, a feature study revealed that stylometric features are better suited to assess topic independent content facets, while for topic oriented content facets Bag-of-Words features serve best. Third, the best features have successfully been used to classify traditional and social media content in both types of content facets. To address the problem of lacking training data in classification, this thesis investigated whether available classification schemes from traditional media can be mapped onto blogs. The experiments revealed that traditional media correlate content wise with selected blogs; therefore, content facets from traditional media can be applied to blogs. Several proposed content facets have successfully been implemented in APA Labs, a Web-based framework for faceted search in traditional and social media. Consequently, APA Labs supports media consumers to navigate and analyze traditional and social media content. This enables any media consumer to search for information according to their personal information need. This is a substantial improvement to individualize search in media.

Zusammenfassung

In traditionellen Medien wird täglich eine grosse Menge an Inhalten publiziert. Zusätzlich wurden soziale Medien in den letzten Jahren immer bedeutsamer, da Menschen Kommentare und Meinungen zu aktuellen Ereignissen in Blogs teilen. Die Qualität dieser Inhalte ist jedoch fraglich, da Blogs weder geprüft noch redigiert werden. Aus der Sicht der Medienkonsumenten ist die Navigation dieser Fülle von Information und das Finden von Inhalten, die den Qualitätsansprüchen genügen, eine Herausforderung. Diese Herausforderung motiviert diese Dissertation: Benutzer sollen auf Basis von Facetten dabei unterstützt werden, Medieninhalte thematischen und qualitätsspezifischen Informationsbedürfnissen entsprechend zu filtern.

Diese Dissertation schlägt hierfür zwei Typen von Facetten vor: (i) themenorientierte, und (ii) themenunabhängige, qualitätsbezogene Facetten. Bezogen auf die Anwendungsdomäne werden für jeden Facettentyp konkrete Facetten vorgeschlagen. Für jede Facette wird eine formale Definition eingeführt. Eine Merkmalsanalyse zeigt, dass stylometrische Merkmale besser geeignet sind um themenunabhängige Facetten zu erhalten während für themenorientierte Facetten Bag-of-Words Merkmale empfohlen werden. Basierend auf den besten Merkmalen werden traditionelle und soziale Medieninhalte in die vorgeschlagenen Facetten klassifiziert. Um das Problem fehlender Trainingsdaten für Klassifikation zu addressieren wird untersucht, ob verfügbare Klassifikationsschemata aus traditionellen Medien auf Blogs angewandt werden können. Ergebnisse zeigen, dass traditionelle Medien und ausgewählte Blogs inhaltlich korrellieren; daher können jene Facetten domänenübergreifend auf Blogs angewandt werden. Einige vorgeschlagene Facetten wurden erfolgreich in APA Labs implementiert, einer Web Anwendung zur facettierten Suche in traditionellen und sozialen Medien. Folglich unterstützt APA Labs Medienkonsumenten bei der Navigation und Analyse von traditionellen und sozialen Medieninhalten. Dies stellt eine substanzielle Verbesserung von Suche in Medien dar.

Acknowledgements

I would like to thank the people who contributed to and supported me in the realization of this work: Prof. Dr. Stefanie Lindstaedt, my professor and advisor. Thank you for critical and constructive discussions about my ideas and concepts. Prof. Dr. Harald Kosch, my second advisor. Thank you for taking the role of my external advisor and examiner. Dr. Michael Granitzer, scientific director of the Know-Center and my division manager for the first three years who always had time for me to discuss my work and my ideas. Thank you for giving me the opportunity and freedom to pursue this thesis. My division manager Wolfgang Kienreich. All the companies with which I conducted challenging projects. Thank you. My former colleague Dipl. Ing. Andreas Juffinger who supported and encouraged me and with whom I had great discussion about research and life. Working with you was a great inspiration and pleasure. Special thanks go to my colleague Dr. Barbara Kump for proof reading this thesis. All other colleagues at the Know-Center. Thank you. My mother, Hilda, who taught me that curiosity is a quality and that you can reach everything if you are committed. You have supported me in learning and advancing throughout my life. My sisters Hilda, Daniela, and Maria. Thank you. My friends for their patience and understanding. And Georg - Thank you for being there for me.

"We are drowning in information but starved for knowledge." (John Naisbitt)

Contents

List of Figures

List of Tables

Chapter 1

Introduction

"Blogging is bringing new voices to the online world" [Lenhart and Fox, 2006].

In the past, news content has been produced mostly by journalists (experts) while traditional media agencies have been responsible for publishing and archiving content. The traditional media agencies have guaranteed a certain level of quality and a journalistic value because the content is typically edited and reviewed before it is published.

The advent of Web 2.0 in 2003 has revolutionized the media domain: As Tim O'Reilly and Dale Dougherty stated in their initial definition of the Web 2.0, *publishing* in Web 1.0 translates into *participation* in Web 2.0[1]. More specifically, in Web 2.0, users have become producers in addition to being consumers of content. This so called *user generated content* is nowadays created by both non experts and experts who share their ideas, thoughts, and experiences in Web 2.0 platforms like Wikipedia[2], Youtube[3], Flickr[4] or blogs.

From all Web 2.0 platforms, especially blogs reflect a cross-section of the public interest consisting of objective reports, opinions, gossip, or personal diaries since blogs are open to almost anyone. Besides, blogs are much more up-to-date: For instance, breaking news and events are often posted in blogs via microblogging platforms like Twitter before they are published by any media agency. Lenhard and

[1] `http://oreilly.com/web2/archive/what-is-web-20.html`
[2] `www.wikipedia.org`
[3] `www.youtube.com`
[4] `www.flickr.com`

Fox investigated the role of bloggers and revealed that 34 percent categorize their blog as a form of journalism [Lenhart and Fox, 2006]. This was also reflected by Mossberg in [Mossberg, 2003] who stated that *"Blogs are in some ways a new form of journalism, open to anyone who can establish and maintain a Web site"*. From this, it can be concluded that at least a subset of blogs is a valuable resource to gather background information to news stories, or to analyze the public feedback towards products, brands, or campaigns.

The amount of information available in the totality of blogs, the so called *blogosphere*, is huge and still expanding. According to eMarketer, a market research company specialized on reports about the Internet, in 2009, around 23.9 million people in the US participated in blogging and by 2013 this number is expected to grow to 32.1 million people[5]. Consequently, blog consumers face the problem of *information overload* [Edmunds and Morris, 2000] which means that there is too much information available to make a decision [Toffler, 1984].

Naturally, the blogosphere contains different aspects or *facets*: from objective reports to opinionated blog posts or personal diaries. Clearly, for blog consumers it is challenging to discover relevant and useful content in this haystack of information available in the blogosphere [Savage, 2010]. In this context, the definition of what is relevant and useful clearly depends on the personal information need of blog consumers. For instance, one blog consumer might be interested in a subjective eye witness account on the Haiti earthquake while another wants to read an objective piece about Haiti. Consequently, a categorization of blogs according to genre or topic and the discovery of content facets useful for an individual's personal information need is crucial to facilitate the navigation and retrieval of blogs.

1.1 Research Questions

The goal of this dissertation research is to classify traditional media content, online news, and social media content into different facets. Such facets can support users to filter content according to the users' personal information need.

Usually, supervised text classification algorithms are used to assign content to a set of categories [Ikeda et al., 2008]. However, supervised text classification algorithms need labeled training data [Chan et al., 2007] to construct a classifier model

[5]http://mashable.com/2009/02/19/user-generated-content-growth/

that reflects the content's most characteristic features [Sebastiani, 2002]. For blog classification, a naive approach would be to directly exploit available categories to which blog posts were assigned to by their creators. However, these categories usually neither stem from a controlled vocabulary nor are commonly agreed upon.

Another idea would be to exploit blog tagging information[6], if available (some blogging systems allow users to assign free-form tags). However, the tagging vocabulary is mostly highly user specific [Sen et al., 2006]. So for a general, non user specific blog categorization, this tagging information cannot be exploited.

In contrast to that, in traditional news media, newspaper editors hence experts categorize news articles into a set of well defined and commonly agreed upon categories like the classical newspaper categories politics, culture, economy, sports, local. Usually these categories follow established standards like for instance the *Colon Classification* [Ranganathan, 1933], a standard for faceted classification of news articles and journals. Consequently, a natural approach would be to investigate whether established categories from newspapers can be utilized to also categorize blogs. This leads to the first research question of this thesis.

> *Research Question 1: "How effectively can classification schemes from traditional media be mapped onto blogs? "*

The quality and credibility of information contained in blogs may be questionable [Mishne and de Rijke, 2006]. Due to the variety of aspects and opinions contained in blogs, it is challenging for blog consumers to find blogs of acceptable quality which actually meet their personal information need [Hearst et al., 2008]. Consequently, if facets encoding different aspects like content quality would be available in addition to the content itself, blog consumers would be able to decide beforehand whether the blog meets their personal information need and quality demands. This leads to the second research question.

> *Research Question 2: "Into which types of content facets can blogs be categorized?"*

In this context, the selection of features for different types of content facets is crucial. Generally, in this dissertation research, content facets are divided into two cat-

[6] According to Wikipedia, a tag is a non-hierarchical keyword assigned to a piece of information or resource

egories: (i) topic oriented facets like blog genre or blog topic, and (ii) topic independent facets which encode measures for information quality as for example the level of subjectivity, emotionality or the trustworthiness of content. Note that literature also describes other content facets like location [English et al., 2002b], time or the role of content (e.g. the content is specifically useful for journalists) [Hearst, 2005]. This dissertation research yet focuses on only the in (i) and (ii) mentioned types of content facets. Hearst also outlines that facets can be hierarchical, in case of this thesis, I focus on flat facets.

Since traditional text mining methods with classical word based models are inherently topic driven [Lex et al., 2010c], in case of the topic independent facets, other features have to be identified. Therefore, this dissertation research tackles the following Research Question three:

> *Research Question 3: "Which types of features are most suitable for detecting topic oriented facets and topic independent facets?"*

The availability of different content facets can also be beneficial in other settings than blog categorization: For example, facets denoting the content quality can foster the automation of Web archival processes or the triggering of crawling processes because best quality resources could be prioritized. Besides, the availability of genre facets could facilitate to filter out low quality content like spam or adult content. This leads to the fourth research question:

> *Research Question 4: "If content facets and features can be identified and extracted from media content, can they be generalized to Web content?"*

1.2 Structure of this thesis

This thesis is structured as follows: In Chapter 2, the technical background of this dissertation research is outlined. This chapter leads step-by-step into the core research areas of this thesis, namely Information Retrieval, Faceted Search, and Text Classification. This zoom view from the topic Information Retrieval to Content Facet Assessment is depicted in Fig 1.1. Section 2.4.2 presents the Text Classification algorithms used within this dissertation research. Chapter 3 introduces a news retrieval framework that has been developed in the course of this dissertation and

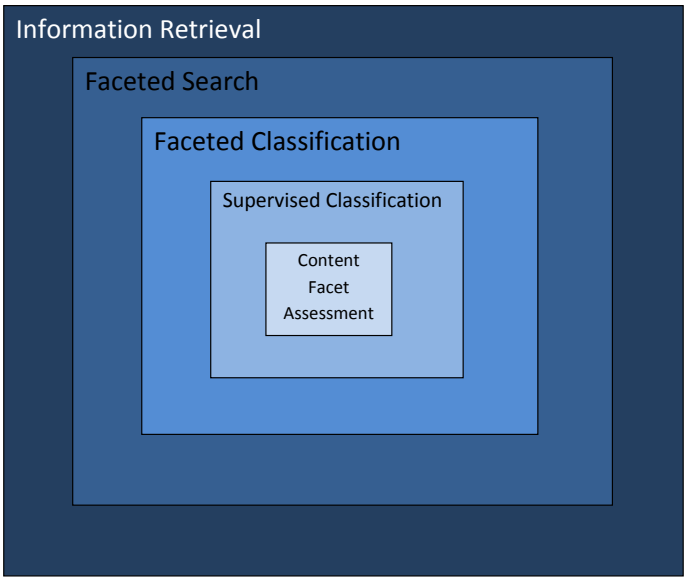


Figure 1.1: The research field to which this thesis contributes can be narrowed down from the general research field of Information Retrieval to the research area of content facet classification

that served as environment for content facet assessment in news and blogs. Chapter 4 outlines the state of the art in content facet extraction and derives the content facets that have been assessed within this thesis. This chapter provides an answer to Research Question 2. In Chapter 5 the conducted content facet experiments in the media domain are outlined and the achieved results are presented and discussed. This chapter firstly answers Research Question 3 and then Research Question 1. In Chapter 6, the insights gained in the content facet experiments are applied to an information quality task in which I participated in the context of the ECML/PKDD Discovery Challenge 2010. This chapter provides an answer to Research Question 4. The thesis concludes in Chapter 7 with a summary of the most important findings and a self assessment on the answering of the proposed research questions.

1.3 Scientific Contributions

In this section, the scientific contributions of this dissertation research are outlined and the own publications are listed.

Chapter 3, Chapter 4, and Chapter 5 present the central own scientific contributions and achievements. The achievements are based on research carried out be

me in the field of news retrieval, blog analysis, text classification, and information quality assessment, in collaboration mostly with Andreas Juffinger, Michael Granitzer, Christin Seifert, Markus Muhr, and Wolfgang Kienreich. The research for this dissertation was performed as part of my work at the Know-Center GmbH[7], Austria's Competence Center for Knowledge Management.

Parts of this dissertation research have already been published in the following 13 publications:

(1) **Elisabeth Lex**. Content Based Quality And Credibility Assessment For Blogs. In *Doctoral Paper at IADIS International Conference on WWW/Internet*, 2009

This work presents an early description of some of the core concepts of the proposed dissertation research. The primary goal of this PhD thesis was to define and derivate a methodology to assess information quality and credibility in news-related blogs. The derived methodology should support the community of blog readers who are interested in current events of public interest to identify blogs worth reading to them. The actual proposal is more general than the first proposal. Now, not a complete methodology is derived but a set of facets is assessed to support users in their personal information need since the definition of quality and credibility definitely lies in the eye of the beholder.

(2) Andreas Juffinger, and **Elisabeth Lex**. Crosslanguage Blog Mining and Trend Visualisation. In *Proceedings of the 18th World Wide Web Conference*, 2009

This paper gives an introduction into blogging and the blogosphere. Some of the motivations for this dissertation thesis (see Chapter 1) were first outlined in this publication. This paper describes a whole blog analysis unit consisting of a cross-language blog crawler, a cross-language blog search unit and a blog visualization module. Note that the crawler has been the work of my colleague Andreas Juffinger et al. [Juffinger et al., 2009b].

Generally, this paper has been a first approach towards a deeper analysis of blogs especially in the news media domain. The finding that there is a correlation between news articles and selected blogs for different search query terms

[7] `www.know-center.at`

and languages has motivated further experiments to map news properties to the blogosphere (see Chapter 5.4.1). In the course of this first approach towards a deeper understanding of blogs, the above mentioned blog visualization module has been designed in form of a *Blog Trend Visualization*. This Blog Trend Visualization shows blog posts to a query over time in comparison to news articles to the query over time. The description of the visualization in Chapter 3.3 has been published in this paper.

The Blog Trend Visualization has been created as part of the APA Labs[8] framework. The APA Labs framework has been designed and implemented by Wolfgang Kienreich and me in the context of a research project for the Austrian Press Agency APA. The main intention behind the project was to foster the retrieval and analysis of German news articles. The description of the APA Labs framework in Chapter 3 has been published in the following publications:

(3) **Elisabeth Lex**, Christin Seifert, Wolfgang Kienreich, and Michael Granitzer. A generic framework for visualizing the news article domain and its application to real-world data. In *Journal of Digital Information Management*, 2009

(4) Wolfgang Kienreich, **Elisabeth Lex**, and Christin Seifert. APA Labs: An Experimental Web-Based Platform for the Retrieval and Analysis of News Articles. In *Proceedings of the first International Conference on the Applications and Digital Information and Web Technologies (ICADIWT08)*, 2008

These publications describe the concept behind the APALabs framework, its correlation with core Web 2.0 business models and the technologies used to implement the framework. Apart from the before mentioned Blog Trend Visualization, other visualization modules have been implemented to evaluate novel and innovative knowledge management services and information visualization techniques. Consequently, the publications also outline the other visualization modules and their functionality. The other visualization modules analyze German newspaper articles, and their description in Chapter 3.3 has been taken from these publications.

Section 5.4 describes the first experiments that were conducted to assess topic oriented facets in blogs. For this, blogs were classified into commonly agreed upon

[8] www.apa.at/labs

newspaper categories following a supervised classification strategy. The description of the concept behind supervised classification has been published by Christin Seifert and me in [Seifert and Lex, 2009].

(5) **Elisabeth Lex**, Christin Seifert, Michael Granitzer, and Andreas Juffinger. Efficient Cross-Domain Classification of Weblogs. In *International Journal of Intelligent Computing Research (IJICR), 1(2)*, 2010.

(6) **Elisabeth Lex**, Christin Seifert, Michael Granitzer, and Andreas Juffinger. Automated Blog Classification: A Cross Domain Approach. In *Proceedings of IADIS International Conference WWW/Internet*, 2009.

(7) **Elisabeth Lex**, Christin Seifert, Michael Granitzer, and Andreas Juffinger. Cross-Domain Classification: Trade-Off between Complexity and Accuracy. In *Proceedings of the 4th International Conference for Internet Technology and Secured Transactions (ICITST)*, 2009.

The description of experiments in Section 5.4.2 are strongly based on these publications. In these publications, especially the problem of lacking labeled training data for supervised blog classification is tackled. This problem is explained in these publications and as a solution, we suggest to exploit news paper categories from German news articles to label unseen German blogs. The description of the related work in this context in Chapter 5 is based on the related work on cross domain classification described in these publications. Besides, the use of a highly efficient centroid based algorithm, the Class Feature Centroid classifier is proposed in this work. The description of the Class Feature Centroid classifier in Chapter 2.4.3 is taken from the algorithm descriptions in these publications.

Chapter 5 outlines the experiments and results achieved in the field of facet classification. This chapter basically contains two major contributions: (i) the results of several feature engineering experiments to identify the best performing content based features for facet classification. And (ii), the results achieved with content based features in a public challenge, namely the TREC challenge 2009.

(8) **Elisabeth Lex**, Michael Granitzer, Markus Muhr, and Andreas Juffinger. Stylometric Features for Emotion Level Classification in News Related Blogs. In *Proceedings of the 9th ACM RIAO Conference*, 2009

In this work, we propose a set of style based, so called *stylometric* features to assess whether a blog post exhibits emotion or not. This publication corresponds closely to Section 5.5.3 but misses a detailed description of the stylometric features.

(9) **Elisabeth Lex**, Andreas Juffinger, and Michael Granitzer. A Comparison of Stylometric and Lexical Features for Web Genre Classification and Emotion Classification in Blogs. In *IEEE Computer Society: 7th International Workshop on Text-based Information Retrieval in Proceedings of 21th International Conference on Database and Expert Systems Applications (DEXA 10), 2010*

In this work, the proposed stylometric features are evaluated and compared to a set of standard text classification features, so called *lexical features*, in terms of classification performance. The description of the stylometric and lexical features in Chapter 5.3.2 and Chapter 5.3.1 is taken from this publication.

(10) **Elisabeth Lex**, Andreas Juffinger, and Michael Granitzer. Objectivity Classification in Online Media. In *21st ACM SIGWEB Conference on Hypertext and Hypermedia (HT2010)*, 2010

Chapter 5, Section 5.5.1 investigates the topic independency of both lexical and stylometric features. This section corresponds closely to this publication in which we show that classifiers trained on lexical features inherently learn topics while classifiers trained on stylometric features are topic independent. This dissertation research thus adds to the state of the art because even though stylometric features have been used in other text classification settings, to the best of my knowledge, their topic independency has not been shown so far.

(11) **Elisabeth Lex**, Michael Granitzer, and Andreas Juffinger. Facet Classification of Blogs: Know-Center at the TREC 2009 Blog Distillation Task. In *Proceedings of the 18th Text REtrieval Conference*, 2009.

In Section 5.5.5, our efforts are described that were carried out within the TREC 2009 blog track. We specifically participated in the blog distillation task which aimed to assess facets in blogs that address quality aspects of blogs. This publication contains a short description of the challenge tasks and outlines our approach and the results we achieved. Section 5.5.5 is an extension of this publication.

(12) Andreas Juffinger, Michael Granitzer, and **Elisabeth Lex**. Blog Credibility Ranking by Exploiting Verified Content. In *Proceedings of the 3rd Workshop on Information Credibility on the Web at 18th World Wide Web Conference*, 2009.

In Section 5.5.4, another topic independent content facet, namely *credibility* is assessed. This paper introduces the concept credibility and proposes an approach to rank blogs by credibility. The experiments described in Section 5.5.4 are an extension of this paper.

(13) **Elisabeth Lex**, Inayat Khan, Horst Bischof, and Michael Granitzer. Assessing the Quality of Web Content. In *Proceedings of the ECML/PKDD Discovery Challenge* 2010.

Chapter 6 provides a generalization of content facets to arbitrary Web content. Besides, several content facets are combined to derive a single content quality score. Section 6.2 describes our efforts carried out for the ECML/PKDD Discovery Challenge 2010. The goal of the challenge was to assess the quality of multilingual Web hosts. For the challenge, the term quality was defined by an aggregate score composed out of Web genre (Web Spam, News/Editorial, Educational/Research, Discussion, and Personal/Leisure) and a set of content facets (trustworthiness, bias, and neutrality). The participation in the challenge resulted in the runner up position. This chapter is an extension of this publication whereas the connection of the challenge to this dissertation research is explained.

To sum up, the first contribution of this thesis is the APA Labs framework, a Web-based application for visually supported faceted search. The APA Labs framework has been introduced and described in the publications (2), (3), and (4) in the above listing.

The second contribution of this dissertation is that it has been shown that classification schemes from traditional media can be mapped onto news related blogs very efficiently (Research Question 1). This has been investigated in the publications (5), (6), and (7) in the above listing.

The third contribution is a classification of content facets into two types, namely topic oriented content facets, and topic independent content facets (Research Question 2). This classification is based on a literature review given in Chapter 4. As a

result, methods to assess both topic oriented and topic independent content facets in traditional as well as social media are suggested and successfully applied. The experiments conducted in this context have been published in the papers (1), (8), and (12) in the above listing.

In Chapter 5, each assessed content facet is defined formally; to the best of my knowledge, such formal definitions are not available and therefore they represent an own scientific contribution.

The fourth contribution is a feature study that compares textual features capturing a document's style with classic bag-of-words features. It has been shown that stylometric features are better suited to assess topic independent content facets while bag-of-words features are better suited for topic oriented content facets (Research Question 3). The proposed stylometric and bag-of-words features have been applied and evaluated in the publications (9), (10), and (11) in the above listing.

A selection of the proposed content facets, namely genre and content quality aspects have been assessed in multilingual Web content in the context of the ECML Discovery Challenge 2010 (Research Question 4). The major contribution here is that with topic independent features, Web genre and content quality has been assessed across different indoeuropean languages (1). This has been published in paper (13) in the above listing.

1.4 Terms and Definitions

Web 2.0 and User Generated Content

Web 2.0 is also referred to as the participatory Web [Decrem, 2006] since in Web 2.0, people not only consume content but also produce content [Hass et al., 2008, Anderson, 2006]. This so called *user generated data* or *consumer generated media* has been arisen during 2005 in the context of Web publishing [Agichtein et al., 2008]. The importance of user generated content is underlined by the fact that the Time magazine voted the participants of Web 2.0 user generated content platforms like Wikipedia, Youtube, Flicker or Blogs as Person of the Year in 2006 [Grossman, 2006].

Blog

A blog represents a chronological ordering of single *blog posts* written by one or a few users, so-called *bloggers* [Macdonald et al., 2010b] A blog post typically covers one major topic whereas the whole blog may tackle a broad range of topics and opinions. Writing a blog is called *blogging*. Blogs may contain textual and/or visual content (images, videos), or simply links to Web pages or other blogs. Blog readers typically have the possibility to comment on blog posts and to add links to external resources in such comments.

Blogosphere

The totality of all blogs and its interconnections is referred to as blogosphere[9].

Blog Analysis

Blog analysis is the analysis of the medium blog. For instance, blog analysis can be used to derive and monitor trends over time from a set of blogs. Generally, blog analysis is often used to identify what people think about a specific topic. As mentioned by Huang et al., blog analysis can be used to measure how the public reacts towards products or brands [Huang et al., 2006a] or to analyze the effects of political campaigns and events on the public [Huang et al., 2006a]. As Huang et al. also outline, via blog analysis the global and personal mood can be assessed.

News Related Blogs

In the context of this dissertation research, news related blogs are defined as blogs that are written by a user or a group of users who are independent of any traditional news agency, and whose primary intention is to comment on current events for some community of relevant size [Lex et al., 2010a].

[9]Definition of Blogosphere in Wiktionary, `http://en.wiktionary.org/wiki/blogosphere`, last accessed May 2011

Facet

The term facet denotes a well defined aspect or a characteristic property of an item. In its original definition, a facet should be mutually exclusive [Taylor, 1992]. Facets can also be regarded as dimensions or feature types [Hearst, 2009b]. Hearst also outlines in [Hearst, 2005] that while facets should by definition be exhaustive and mutually exclusive, in reality this is often not the case.

Content Facet

In this dissertation research, the term content facet is understood that a content facet describes a facet that has been derived *from content only*. Such a content facet denotes genre, topic, or quality related aspects of the content.

Topic Oriented Content Facets

Topic oriented content facets are defined within this dissertation research as facets that correspond to the topic of content.

Topic Independent Content Facets

Topic independent content facets are defined within this dissertation research as facets that are relatively independent of the topic of content and that encode dimensions useful for assessing the quality and credibility of content. Such topic independent content facets are for instance subjectivity, emotionality, and credibility.

Web Classification

In Web classification, the goal is to assign Web sites, or Web pages, or Web hosts to a set of pre-defined categories [Qi and Davison, 2009, An et al., 2004].

Information Quality

Information quality denotes the quality of the content of information. and it depends on how and by whom information is used. Generally spoken, information exhibits quality if it provides value for specific users [Eppler, 2003].

(Un-)structured data

In contrast to structured data which has a clear and machine understandable structure, unstructured data does not exhibit such. A typical example for structured data is a relational database with well defined database entries [Delbru, 2010]. In practice, most data exhibits at least some structure: for instance, most documents feature title and content. On the Web, data is often structured via markup languages (HTML) [Manning et al., 2008].

Information Need

An information need describes the interest of a user. Information needs are often expressed by a means of a query to the system [Losee, 1997], for instance a query to a blog search engine like Technorati[10].

Index

As outlined by Baeza-Yates et al. in the famous book Modern Information Retrieval, an index consists of a collection of words as well as a pointer to where in a document a particular word is located [Baeza-Yates and Ribeiro-Neto, 1999].

Concept

A concept is regarded as a *"unit of knowledge"* [Antia, 2007] whereas a concept typically is an abstract representation of an object with its properties [Antia, 2007].

Metadata

As described by Taylor in [Taylor, 2003], metadata describes the characteristics of a document by means of structured data.

[10] `http://technorati.com`

Chapter 2

Information Retrieval and Text Classification

"If we would have new knowledge, we must get a whole world of new questions."(Susanne Langer)

2.1 Introduction

On the Web, a huge and steadily growing amount of information is available. In May 2010, the EMC Corporation[1] released the results of an EMC study titled "The Digital Universe Decade - Are you Ready?" which measures and predicts the amount of digital information created per year worldwide. In 2009, the amount of digital information grew to 800 billion gigabytes (0.8 Zettabytes[2]). In 2010, this amount has been estimated to reach 1.2 Zettabytes. For the future, the EMC study predicts that the amount of digital information will grow by a factor of 44 from 2009 to 2020 [Farmer, 2010].

Navigating this haystack of information is challenging - especially for Information Retrieval (IR) systems, search engines, respectively [Savage, 2010]. Information needs are changing; as stated by Oren Etzioni[3], director of the Turing Center[4], the

[1]EMC is both developer and provider of information infrastructure technology and solutions

[2]One Zettabyte equals one trillion giga bytes.

[3]Oren Etzioni's home page, `http://www.cs.washington.edu/homes/etzioni/`, last accessed June 2011

[4]Turing Center at University of Washington, `http://turing.cs.washington.edu/`, last accessed June 2011

problem of answering questions is transformed from *"finding a needle in a haystack to a process of being presented with a variety of needles and choosing the one you want"* [Garner, 2007].

This dissertation research addresses this challenge by exploiting concepts from the field of *Faceted Search*. Faceted Search, also called Faceted Navigation or Faceted Browsing, enables users to filter a document collection and to narrow down search results by certain aspects, the facets.

This chapter outlines the theoretical backgrounds for this dissertation. It is divided into three main sections: Firstly, in Section 2.2, the foundations of Information Retrieval are discussed and standard techniques for document representation are outlined. Secondly, this chapter narrows down to Faceted Search and Faceted Classification in Section 2.3. Thirdly, in Section 2.4.1, the basics of Text Classification are outlined and the classification algorithms used within this dissertation research are discussed.

2.2 Information Retrieval (IR)

Information Retrieval (IR) is the task of retrieving documents from a set of unstructured data according to an information need [Manning et al., 2008]. Information Retrieval (IR) is a subfield of Natural Language Processing (NLP)[5] since in IR techniques from NLP are applied [Manning and Schuetze, 2003]. IR enables users to explore a document collection according to their *personal information need*. Typically, the information need is expressed using a *search query* consisting of a set of keywords. Such a search query is entered in the search engine that returns a list of search results. This is called the *"unassisted keyword search"* [Dennis et al., 2002] paradigm[6].

Since usually more than one document matches a query, IR systems need to determine whether a document is actually relevant to the query. In other words, given a search query, IR systems aim to retrieve relevant documents while leaving out non relevant documents [Baeza-Yates and Ribeiro-Neto, 1999]. This actually

[5]Natural Language Processing (NLP) aims at making computers understand human natural languages.

[6]Other search paradigms are e.g. assisted keyword search, directory-based search, and query-by-example. For a review of these paradigms, see [Dennis et al., 2002]

represents the most common IR task, the *ad hoc retrieval task*. Most popular today's search engines as for instance Google[7] basically perform ad hoc retrieval.

2.2.1 History of Information Retrieval

Since the first mention of Information Retrieval in 1945 [Bush, 1945], the primary goal of IR has been to index text and to search for information in a document collection. Research in IR has been pursuit since then, and various IR models have been proposed. The introduction of the Cranfield evaluations in the 1960ies and 1970ies has enabled to experimentally evaluate the applicability of the models and techniques, albeit on relatively small text corpora [Cleverdon et al., 1966, Cleverdon, 1970]. The advent of Web search engines has coined the need for large scale IR systems: The introduction of the - still ongoing - influential Text Retrieval Conference (TREC) in 1992 [Harman, 1992] by the National Institute of Standards and Technology (NIST) from the US Department of Defense fostered research on large scale IR systems since TREC provides both large text corpora as well as a standardized evaluation infrastructure [Singhal, 2001]. The goal of TREC is to boost dissemination and technology transfer in the field of IR. The TREC conferences consist of a set of tracks addressing different types of problems. Each track organizes a challenge for which both data sets as well as a set of problems is given. When the results for the challenge have been submitted, TREC holds workshops with the goal to discuss the different approaches and the results. Besides, at the workshops ideas for future tasks are presented and discussed [Voorhees, 2005].

2.2.2 The Information Retrieval Process

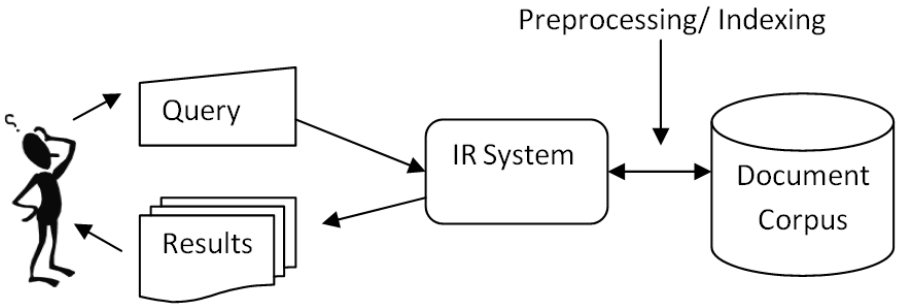
The goal of an Information Retrieval Process is to meet the information need of a user [Baeza-Yates and Ribeiro-Neto, 1999]. Baeza-Yates et al. describe the IR process as follows: As a first step, the document collection relevant for the task has to be selected and the application domain and the goal of the retrieval process have to be defined. Then, the documents in the collection are indexed in the IR system. If an IR system represents a document with all its terms, this is called a *full text* logical view. However, for large document collections, it is more feasible to reduce the number of terms per document to a subset of representative

[7] `www.google.com`

keywords. This can be achieved via several text operations or pre-processing techniques [Baeza-Yates and Ribeiro-Neto, 1999]. For this, a variety of methods exist; the most important are described in Section 2.2.2.

The preprocessed documents are then stored in the IR system in form of a search index. A search index is a datastructure that is used to find items to a query. Having the documents in the index, users can formulate an information need specified by a query. As a result, a subset of the document collection is retrieved that is typically ranked by its relevance to the query. The steps of the IR process are shown in Figure 2.1[8]. In the IR Process, special attention has to be paid to

Figure 2.1: The Information Retrieval Process



the index creation. An established index representation is the *inverted index*. An inverted index consists of a dictionary holding all terms of the document collection and a mapping in which documents and at which position a particular term has occurred. The advantage of the inverted index its efficient representation, since usually, this data structure is highly sparse because many terms occur only in a few documents [Baeza-Yates and Ribeiro-Neto, 1999]. This connection is further outlined in Section 2.2.3.

Document Preprocessing

In natural language, some words carry more semantic information about the content of documents than others [Baeza-Yates and Ribeiro-Neto, 1999]. Zipf's law describes this matter. Zipf's law says that some words occur more often in documents than others. If all terms of a set of documents are sorted by their frequency, the probability of their frequency is indirectly proportional to their position in the

[8]The image is an adaption of `www.cs.huji.ac.il/~dbi/lectures/ir/Introduction-IR.pdf`

frequency table [Manning and Schuetze, 2003]. For instance, in the Brown corpus, the most frequent word, *the*, accounts for 7% of all words in the corpus. Following Zipf's law, 3.5% of all words represent the second most frequent word, *of*[9]. In case of the Brown corpus, the most frequent words apparently have no relevant semantic meaning. It is clearly necessary to identify words with more semantic meaning and to further use only these as index terms [Baeza-Yates and Ribeiro-Neto, 1999]. This can be accomplished using different document preprocessing steps; the most important are outlined in the next paragraphs.

In general, document preprocessing results in a reduction of the index size and consequently, in a higher performance of the IR system. Especially for Web search, performance enhancement is crucial since the amount of searched data is large. As a showcase one can use the recent introduction of Google Instant[10], a search function that dynamically provides search results to a query in the instant the user types the query into the search field.

Lexical analysis of the text The lexical analysis of texts aims at converting a set of coherent characters into a sequence of atomic words that potentially can serve as index terms. This is also referred to as *tokenization*. In its simplest way, tokenization can be achieved by using the whitespace character as delimiter. However, in practice, this is not sufficient. Also punctuations, hyphens, and the case of letters should be considered [Baeza-Yates and Ribeiro-Neto, 1999].

The identification of index terms based on a lexical analysis should be treated with care. For instance, the use of digits as index terms may introduce a strong bias since without context, the information content of e.g. a percent value is often unclear. Also problematic are hyphens since their usage is often inconsistent. For instance, "Part-of-Speech" means the same concept as "Part of Speech". Therefore, some lexical analyzers remove all hyphens anyway. However, this may change the whole semantics of a text. For example, the sentence "a man-eating tiger" denotes an animal that eats humans. The same sentence without hyphen "a man eating tiger" is a man who eats tiger[11]. In general, since there are many exceptions, often semi automatic approaches are feasible [Baeza-Yates and Ribeiro-Neto, 1999].

[9]Brown Corpus, `http://en.wikipedia.org/wiki/Brown_Corpus`, last accessed June 2011

[10]Google Instant, `http://www.google.com/instant`, last accessed on Feb 2011

[11]Hyphens part 1: avoiding confusion, `http://dbennison.wordpress.com/2010/02/20/hyphens-part-1-avoiding-confusion/`, last accessed June 2011

Elimination of stop words Stop words are terms that occur very often in a document. Therefore their ability to discriminate between documents is rather low. Usually, stop words have rather a syntactic function than particular relevance for the document's content. By removing such stop words, the size of possible index terms per document is strongly decreased. Examples for English stopwords are conjunctions ("and", "or", etc.) or prepositions ("in", "to", etc.), or articles ("the", "it",..) [Baeza-Yates and Ribeiro-Neto, 1999]. On the Web, a large number of stop word lists for different languages is available[12].

Stemming Stemming reduces words to their stem or root form, respectively. In other words, a word stem is the fraction of a word which is left after the removal of its affixes, i.e. its prefixes or suffixes. For instance, the term "books" is stemmed to the singular form "book" [Baeza-Yates and Ribeiro-Neto, 1999]. There exist a manifold of different stemming algorithms - for this dissertation research, the highly flexible Snowball framework for stemming algorithms has been used[13]. The Snowball framework exhibits an implementation of the Porter stemmer that uses a suffix list for removing suffixes.

Part-of-Speech Tagging Part-of-Speech (POS) Tagging aims to assign parts of speech *tags* to each token in natural language text. The standard tag set of a Part-of-Speech Tagger consists of eight POS: adjectives, adverbs, conjunctions, determiners, nouns, prepositions, pronouns, and verbs. Note that interjections and punctuations are also important tags, yet they are not included in standard tag sets. Naturally, POS Taggers are strongly language dependent since different languages have different POS, grammar, and punctuation rules. To tag English texts, often the *Brown Corpus Tag Set* [Greene and G.M, 1971] and the *Penn Treebank Tag Set* [Marcus et al., 1994] are used. The tags extracted by a POS Tagger can further be used to extract more sophisticated entities using Information Extraction and Named Entity Recognition.

Information Extraction Information Extraction aims at extracting *"structured information from unstructured text"* [Konchady, 2008]. More specifically, unstruc-

[12]For example see `ftp://ftp.cs.cornell.edu/pub/smart/english.stopforEnglishstopwords`

[13]Snowball, `http://snowball.tartarus.org/`, last accessed Jan 2011

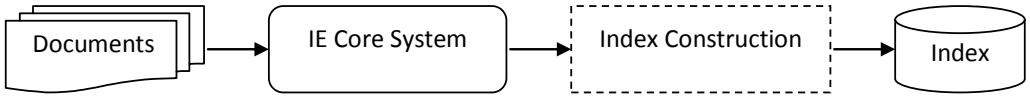
tured text is transformed into structured information by assigning terms a meaning [Konchady, 2008]. Note that a term can consist of multiple tokens. The information to be extracted is typically specified in templates[14]. Information Extraction involves five major subtasks. The first subtask is *Segmentation*, with the goal to identify starting and ending boundaries of text snippets. For instance, a segmentation step can be used to extract the title of a document or the author and the publishing date of a blog. The second subtask, *Classification*, determines the correct label for a text segment. For instance, "Elisabeth Lex" should be assigned to the category "author" of this dissertation. The third step, *Association*, determines which from the extracted text snippets belong together. This step enables to identify relations between extracted text snippets. The fourth step, *Normalization* normalizes information to a standard format in order to foster comparability. For instance, an extracted date may be "07-01-2011" or "7th Jan 2011". Both expressions refer to the same date but without a normalization step, this connection would be impossible to identify. The fifth step, *Deduplication*, is used to reduce redundant information. For instance, an extracted person appears more than once in a blog and consequently, it will be extracted more than once. The deduplication step will guarantee that this record is stored only once. The resulting tokens finally represent the index terms [McCallum, 2005].

Note that a subfield of Information Extraction, *Named Entity Recognition (NER)*, aims to further interpret the extracted entities. For this, entities are classified into more sophisticated and abstract categories like persons, locations, companies, expressions of time, quantities, etc.

As described in [Abramowicz, 2003], Information Extraction and Information Retrieval can be seen as complementary and both techniques can be combined in various ways. For instance, IR could be used within an IE system for retrieving a subset of documents from a document collection whereas the actual IE process is performed on the subset [Robertson and Gaizauskas, 1997]. On the other side, IE can be used in an IR system to identify terms for document indexing. For instance, a document collection can be indexed using Named Entities that have been extracted by an IE system. Figure 2.2 shows the process of IE to identify index terms. Note that the figure has been adopted from [Abramowicz, 2003]. Firstly, the documents are passed to the IE core system that extracts the target entities. These entities are

[14]A template is a user defined structure, e.g. company information

Figure 2.2: The application of IE to identify index terms for IR.



then used as index terms in the IR system. Within this thesis, this is actually the major application of Information Extraction.

2.2.3 Vector Models

In Information Retrieval based on Vector Models, documents and queries are represented by term vectors whereas each term of a document corresponds to a dimension in the vector space. The document representation is typically the classic Bag-of-Words (BoW) representation. In the Bag-of-Words representation, each documents is regarded as a bag of words; a bag of words consists of an unordered collection of terms, whereas grammar and word order are ignored [Salton and Buckley, 1988]. The advantage of the Bag-of-Words model is clearly its simplicity as well as its computational efficiency. Note that yet in some applications like text compression and speech recognition, the word order is important [Wallach, 2006].

The advantage of vector models is that in the vector space, linear algebra methods can be used to compute the similarity between queries and documents and in-between documents because the general assumption is that spatial proximity corresponds to semantic similarity [Errecalde et al., 2008]. In the following, different vector models are outlined.

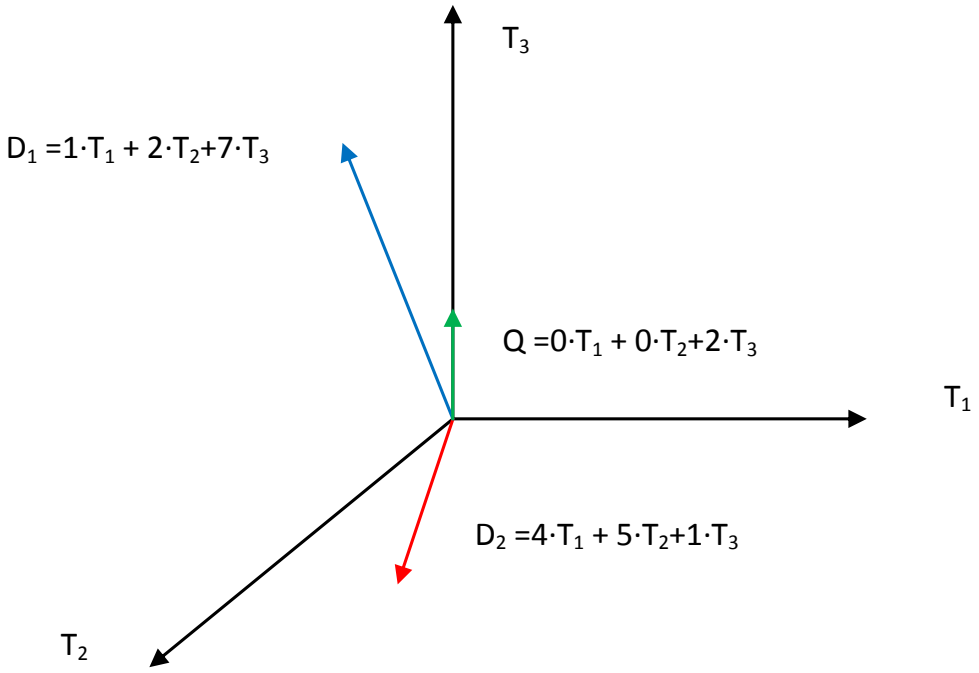
The Boolean Model

In the Boolean model, term vectors encode whether a term is present or absent in a document. More specifically, a term's coordinate is "1" if the particular term is in the document, and "0" otherwise. In the Boolean model, the similarity of documents can be calculated using a geometric measure, e.g. an angle or a vector distance in $\Re$. The Boolean model requires that the queries are formulated as Boolean expressions. The advantage of the Boolean model is clearly its simplicity. However due to its binary relevancy decision ("true" or "false"), it is more data retrieval than information retrieval [Baeza-Yates and Ribeiro-Neto, 1999].

The Vector Space Model

The Vector Space Model (VSM) [Salton, 1971] is a generalization of the Boolean model. Until now, the VSM is one of the most commonly used models in ad-hoc retrieval [Salton et al., 1975]. The advantage of the VSM is its simplicity and the use of spatial proximity for semantic similarity [Manning and Schuetze, 2003]. In the VSM, documents are represented by *term vectors* whereas each dimension in the vector space corresponds to a term. In other words, each vector denotes a document and each term in the vector corresponds to a term in the document. The VSM enables a mathematical computation of e.g. document similarity, distance calculation, addition, subtraction and transformation of documents. Figure 2.3 shows an example of a Vector Space Model with three dimensions, terms T_1, T_2, T_3, respectively. The entities in the vector space are a query vector $q = (0, 0, 2)$ and two document

Figure 2.3: The Vector Space Model)



vectors, $d_1 = (1, 2, 7)$ and $d_2 = (4, 5, 1)$. In this example, the coordinates in these vectors are derived from the frequency of a term in a document. These coordinates are also referred to as *term weights*. The collection of documents then forms the so-called *Document Term Matrix* that contains the frequency of all terms in the

document collection. The rows of the matrix correspond to the documents and the columns to the terms.

In order to improve the vector representation, a more sophisticated weighting can be applied to give a high weight to the most important features. A common weighting scheme is TF-IDF [Salton et al., 1975]. TF-IDF combines the term frequency (TF) and the inverse document frequency (IDF). TF encodes how often a term occurs in a document while IDF is the number of documents the term occurs in: $IDF_i = 1/DF_i$. IDF is a measure for a term's importance for the overall document corpus. A commonly used combination of TF and IDF is TF-IDF [Manning et al., 2008]:

$$TF - IDF_{t,d} = TF_{t,d} \times IDF_t \tag{2.1}$$

This equation can be interpreted so that the more often a term occurs in a document, and the less it occurs in the other documents of the set, the higher its weight.

In Figure 2.3, d_1 has the smallest angle with q; that means that d_1 is more similar to q than d_2 and therefore also more relevant to q. There are different methods to compute the similarity of a document d_i and a query vector q. The most important method, the *Cosine Similarity*, is described in the next section.

Cosine Similarity

In the Vector Space Model, the similarity of documents can be determined by computing the so-called *cosine similarity*[15]. The Cosine Similarity is defined in Equation 2.2.

$$cos(\vec{q}, \vec{d}) = \frac{\sum_{i=1}^{n} q_i d_i}{\sqrt{\sum_{i=1}^{n} {q_i}^2} \sqrt{\sum_{i=1}^{n} {d_i}^2}} \tag{2.2}$$

Note that $\vec{q}$ and $\vec{d}$ are n-dimensional vectors in $\Re$. If $\vec{q}$ and $\vec{d}$ are normalized, the cosine similarity is reduced to a simple dot product [Manning and Schuetze, 2003].

[15]The cosine similarity enables to rank documents by their similarity to the query. This is also referred to as *Relevance Ranking* which is a sub-subject of Information Retrieval.

The cosine similarity is then computed using the dot product and magnitude as

$$cos(\vec{q}, \vec{d}) = \frac{q \cdot d}{||q|| \cdot ||d||} \tag{2.3}$$

Note that the document lengths are normalized. Without length normalization, longer documents would be regarded as being more similar to a query q than shorter documents.

2.2.4 Measures for Evaluating Information Retrieval Systems

The performance of information retrieval systems can be assessed by various performance measures; in the following, five measures are described that are important for this thesis: (i) Precision, (ii) Recall, (iii) f-Measure, (iv) MAP, and (v) NDCG.

The performance measure Precision is defined at the number of returned search results which are actually relevant to a user's information need. The equation for precision is given in Equation 2.4 [Manning et al., 2008]:

$$Precision = \frac{|\{RelevantDocuments\} \cap \{RetrievedDocuments\}|}{RetrievedDocuments} \tag{2.4}$$

Recall is defined as the number of relevant documents returned by the information retrieval system in relation to the number of relevant documents available in the document collection. The equation for precision is given in Equation 2.5 [Manning et al., 2008]:

$$Recall = \frac{|\{RelevantDocuments\} \cap \{RetrievedDocuments\}|}{RelevantDocuments} \tag{2.5}$$

Additionally, other measures have been introduced that combine precision and recall in a way that considers also the relevance ranking of the documents. For instance, the precision can be derived at a particular cutoff, e.g. 5, or 10 documents. A cutoff level of 5 defines the retrieved set as the top five documents in the ranked list [Buckley and Voorhees, 2000]. This measure, denoted as $R - Precision (R - Prec)$ denotes how well a method ranks relevant documents before non relevant documents.

Obviously, there is a tradeoff between precision and recall. For instance, if an IR system returns all documents in the collection, recall is 100% but precision is rather low. If both precision and recall should be considered, the *f-measure* can be used. The f-measure is given in Equation 2.6 [Manning et al., 2008]:

$$f - measure = \frac{1}{\alpha \frac{1}{Precision} + (1-\alpha)\frac{1}{Recall}} \tag{2.6}$$

where α determines the weight of precision and recall [Manning and Schuetze, 2003].

In Information Retrieval, also the *Mean Average Precision (MAP)* is often used to evaluate the relevance ranking of documents. MAP corresponds to the arithmetic mean of the precision values achieved on a document collection for a set of queries. MAP is computed as follows [Manning et al., 2008]:

$$MAP(Q) = \frac{1}{|Q| \sum_{j=1}^{|Q|}} \frac{1}{m_j} \sum_{k=1}^{m_j} Precision(R_{jk}) \tag{2.7}$$

where R_{jk} corresponds to a set of ranked Information Retrieval results starting from the top results until the relevant document d_k is retrieved and Q denotes a set of queries. When a relevant document d_k is not retrieved, the corresponding MAP value is 0 [Manning et al., 2008].

Another Information Retrieval measure that is relevant for this dissertation research is the *Normalized Cumulative Gain (NDCG)* [Järvelin and Kekäläinen, 2000]. NDCG consists of the *Discounted Cumulative Gain (DCG)* which indicates the usefulness (gain) of a document based on its position in the ranked search result list. The rationale behind this measure is that relevant documents should have a high rank position in a search result list. NDCG is therefore computed as follows:

$$NDCG = \frac{DCG}{IdealDCG} \tag{2.8}$$

2.2.5 Application Fields of Information Retrieval Systems

As stated in [Feldman, 2004], it is difficult to obtain relevant and useful information when it is needed, even though enough information is available. This phenomenon is referred to as *Information Overload*. Information retrieval systems address this phenomenon by providing means to efficiently retrieve textual infor-

mation [Crestani, 1998] and by enabling users to browse and filter large document collections [Manning et al., 2008].

According to [Manning et al., 2008], there are three major application fields of IR systems[16].

Personal information retrieval

Personal IR differs from standard IR so that in personal IR, the piece of information to be retrieved is already known and has possibly already been processed by the user. In personal IR, users aim to retrieve previously created content that they need again for some reason [Boström, 2006]. For example, writing a presentation may require searching the Web, but it also involves gathering information from existing information sources like documents, spreadsheets, or E-Mail located on one's computer [Dumais et al., 2003]. To address this information need, operating system providers have directly integrated IR techniques into their systems. For instance, Apple's Mac OS X provides the desktop search Spotlight[17] and Windows 7's Windows Search[18] that can be used to locate files, E-Mail messages, and other items. Besides, most E-Mail programs use Information Retrieval techniques to filter out Spam E-Mails and to automatically sort out E-Mails into user defined subfolders based on e.g. the sender information. There are two main challenges for personal IR. Firstly, there are many different types of documents available on a computer and secondly, the IR methods need to be highly efficient in respect to processing time and the available resources [Manning et al., 2008].

Enterprise Information Retrieval

Enterprise Information Retrieval is the task of searching and filtering document collections from a company's knowledge base. Typically, the knowledge base is stored on an internal centralized file system. The IR system has to be able to search over the whole knowledge base [Manning et al., 2008]. Enterprise IR is also related to domain specific retrieval. An example for domain specific retrieval is patent retrieval where

[16]Nowadays, also the applications Social Search and Mobile Search can be regarded as novel and important IR applications

[17]Apple Spotlight, `http://www.apple.com/at/macosx/what-is-macosx/spotlight.html`, last accessed in Feb 2011

[18]Windows Search, `http://www.microsoft.com/windows/windows-7/features/windows-search.aspx`, last accessed in Feb 2011

the goal is to search within a collection of patents [Atzmüller and Landl, 2009]. An example of an open source enterprise solution is Apache Solr[19].

Web Search

Web search is the task of searching the huge amount of documents available on the World Wide Web. Since the documents are stored on millions of computers worldwide, the main challenges are to gather documents for indexing and to create efficient IR systems [Manning et al., 2008]. Since its introduction, Web search exhibits two search paradigms[20]: (i) navigational search and (ii) direct search. In navigational search, a hierarchical structure is used that enables users to browse search results and to narrow down the result set in a predetermined order. Popular examples for navigational search are Yahoo! Directory[21], and the Open Directory Project DMOZ[22]. In direct search, users just express their search queries in form of a set of keywords entered in a search field. The currently most popular Web search engines Google, Bing[23], Yahoo! Search[24] operate this way [Tunkelang, 2009].

In direct search, a user simply enters a search query and retrieves a set of search results that are ranked by relevance. However, as more and more unstructured data is available on the Web, the need for more sophisticated search techniques has been coined and addressed by the field of *exploratory search*. According to [White and Roth, 2009], exploratory search is more a process than a single search task. As stated in [Marchionini, 2006], human intelligence should be more actively integrated into the search process. Based on these considerations, Marchionini established a whole new area of research: Human-Computer Information Retrieval (HCIR). As Marchionini explains, HCIR is a hybrid approach that combines ideas from human-computer interaction with interactive user interfaces. The goal of HCIR is to enable users to influence the outcome of the search process by exploring the data set by means of user interaction [Marchionini, 2006].

Exploratory search and HCIR foster a deeper understanding of the search topic

[19] Apache Solr, `http://lucene.apache.org/solr/`, last accessed in Feb 2011

[20] The information about these two search paradigms is based on the Call for Participation of the First SIGIR 2006 Workshop on Faceted Search, `http://sites.google.com/site/facetedsearch`

[21] Yahoo! Directory, `http://dir.yahoo.com/`, last accessed Feb 2011

[22] DMOZ, `http://www.dmoz.org/`, last accessed Feb 2011

[23] `http://www.bing.com`

[24] `http://www.yahoo.com/`

itself as well as of the document collection. One of the key technologies in exploratory search is *Faceted Search*. Faceted search combines the advantages of both direct as well as navigational search[25]. More specifically, faceted search enables to firstly retrieve results from a document collection using direct search and secondly, to navigate the search result set by different aspects and facets. Meanwhile, faceted search has become a popular user interaction mechanism in E-commerce sites as for example Amazon[26] and eBay[27] [Broder and Maarek, 2006]. Faceted search is described in detail in Section 2.3.

2.3 Faceted Search and Faceted Classification

Faceted search, also referred to as faceted navigation or faceted browsing, is a technique that enables the user to explore and narrow down a collection by filtering the available information according to different aspects of information, the facets [Tunkelang, 2009]. As in direct search, a user starts by formulating a search query. Given this query, the Information Retrieval system searches the document collection and a list of search results is retrieved that is ranked by relevance. From the search results, different facets can be compiled that enable the user to navigate the search results [Tunkelang, 2006].

Faceted search has become an integral technique in the field of exploratory search [Zelevinsky, 2010]. Faceted search systems exploit faceted classification to enable users to navigate content according to their personal information need. Generally, faceted classification

> *"decomposes compound subjects into foci in component facets, offering expressive power and flexibility through the independence of the facets"* [Tunkelang, 2009]

In other words, a faceted classification system allows to assign different facets to the same object. The term facet should not be confused with the term category: the difference here is that items are placed into one or more categories while mul-

[25]This has also been mentioned in the Call for Participation of the First SIGIR 2006 Workshop on Faceted Search, `http://sites.google.com/site/facetedsearch`

[26]`http://www.amazon.com`

[27]`http://www.ebay.com`

tiple facets can be assigned to items [Hearst, 2009a]. Generally spoken, each facet typically captures different characteristics of the same resource.

> **Example:** A blog typically exhibits, among others, the following characteristics: author information, blog topic, published date, etc. These characteristics can be used in a faceted search system: users can then filter blogs by an author facet, a topic facet, or a date facet.

Consequently, faceted classification systems are, among others, very useful to organize and navigate social media and blogs. **The ideas of faceted classification have significant impact on this dissertation research since faceted classification is used to address the personal information need of users.**

Meanwhile, many popular Web sites and E-commerce providers like eBay or Amazon provide faceted search in their interfaces. Also, software vendors like Endeca and Mercado sell faceted search applications [Ben-Yitzhak et al., 2008].

The feasibility of faceted search and search interfaces with faceted search has been investigated by the research group around Marti Hearst[28] from the University of California in Berkley. In [English et al., 2002a], they conducted a user study on a proposed search interface, the FLexible information Access using MEtadata in Novel COmbinations (Flamenco)[29] search interface framework. The Flamenco search interface[30] exploits faceted hierarchical metadata to provide access to a large multimedia collection. In Flamenco, hierarchical faceted metadata can be used to both narrow as well as expand search queries. In a user study with 19 participants they found that the system has been well adopted by the participants of the study [English et al., 2002a]. In later work, Hearst also suggested design recommendations for such faceted search systems [Hearst, 2006a] based on the insights gained with the Flamenco project.

2.3.1 Facet Extraction

The identification of suitable facets strongly depends on the application domain. In the news domain, the facets "topic", "newspaper category (genre)", as well as facets

[28]Marti A. Hearst, `http://people.ischool.berkeley.edu/~hearst/`, last accessed in Feb 2011

[29]Flamenco stands for FLexible information Access using MEtadata in Novel COmbinations

[30]The Flamenco Search Interface Project, `http://flamenco.berkeley.edu/index.html`, last accessed Feb 2011

denoting quality aspects like "objective", "credible", "emotional" may support the user to (i) navigate news articles by topic and/or genre, and to (ii) filter content in respect to their information need.

> **Example:** A user might be interested in an emotional eye witness account about the Haiti earthquake while another user is looking for some objective facts. The advantage of facets is clearly that they facilitate exploring a collection according to a user's personal information need.

The extraction or creation of meaningful facets requires not only knowledge about the application domain but also about the document collection of interest. In order to reduce information overload, facets should be preferred that cover a sufficiently large number of search results. This can be computed via a greedy solution where at each step, the facet with the maximum number of unseen documents is selected [Serdyukov, 2010, Liberman and Lempel, 2009].

2.3.2 Facet Type Classification

According to [Tunkelang, 2009], facets can be roughly divided into three groups whereas the differentiation is based on the facets' values: (i) nominal facets, (ii) hierarchical facets, and (iii) numerical facets. However, since Tunkelang specifically concentrates on the resulting query metaphors, this classification is not applicable for this dissertation. In this thesis, the focus is on providing facets for unstructured textual content. Therefore, another categorization of facets derived from unstructured content is needed.

In a tutorial on Faceted Search held at the World Wide Web conference (WWW) 2010 [Serdyukov, 2010], three approaches have been described to derive facets from either metadata or content. The focus of the tutorial has been on Enterprise and Desktop search; however, the description of how to derive facets corroborates some findings of this dissertation research. Therefore, the approaches are described in the next sections.

Facets derived from structured metadata

The use of structured metadata is a very common approach towards faceted search. Users can narrow the search result by navigating metadata fields and values. Struc-

tured metadata is typically stored in a metadata record that contains well-defined elements. Each metadata record contains a limited number of pre-defined elements that represent specific attributes of a resource, e.g. the name of each element, and the meaning of each element [Taylor, 2003].

> **Example:** A metadata record in the ACM Digital Library[31] or in the Web based citation system DBLP[32] consists of keywords, the title, the authors, the conference, and the publisher of a scientific paper. Both the ACM Digital Library as well as DBLB enable users to filter scientific publications by the given metadata records [Hearst and Stoica, 2009].

The advantage of this approach is that the metadata is properly defined and consequently well understood. However, especially for online media and Web 2.0 based social media, standardized metadata is lacking. Most metadata in Web 2.0 consists of tagging information or labels that have been assigned to resources by users. The vocabulary of these tags and labels is typically rather heterogeneous, diverse and not commonly agreed upon [Farooq et al., 2007]. This challenge is addressed in the next section.

Facets derived from unstructured metadata

On the Web, a huge amount of unstructured metadata is available in form of tagging information or user generated labels. Tagging is a Web 2.0 technology that enables users to annotate Web resources with arbitrary keywords [Abel et al., 2010]. Tagging has been exploited in the past to assign blogs to a set of predefined categories [Sun et al., 2007].

As mentioned earlier, the disadvantage in tagging is that tags are typically not categorized and that the tagging vocabulary is rather heterogenous. This is referred to as *"vocabulary problem"* [Furnas et al., 1987]. The vocabulary problem denotes that one user might assign different keywords to a resource than the keywords used by other search users [Tunkelang, 2009].

> **Example:** A person might tag a news related blog with *news* while

[31] ACM Digital Library, `http://portal.acm.org`, last accessed Feb 2011

[32] The DBLP Computer Science Bibliography, `http://www.sigmod.org/dblp/db/index.html`, last accessed Feb 2011

another might tag a personal blog describing news about her family also with *news*.

Recent work identified two types of taggers, namely describers and categorizers [Koerner, 2009], who tag resources quite differently. The impact of the different tagging behavior is one of the reasons for the heterogeneous tagging vocabulary. A solution to this is to guess the tag purpose and the tag meaning [Koerner et al., 2010] and to find relevant tags. Another challenge is that usually tagging vocabulary is rather subjective, not consistent across blogs, and may change over time.

Due to all these drawbacks, it is therefore generally not possible to directly use individual tagging information, respectively folksonomies[33]. If tagging information should be used in a faceted search scenario, first, the tagging information needs to be structured [Serdyukov, 2010]. For instance, a mapping from an evolving folksonomy to pre-defined and commonly agreed upon categories can be learned via supervised classification [Mathes, 2004].

A common approach to using tags in a faceted search system are so-called clickable *tag clouds* [Börkur and van Zwol Roelof, 2010]. A tag cloud shows a set of tags whereas the font size of the tags denotes the number of resources the tag has been applied to. Tag clouds are common Web 2.0 tools that have been integrated into popular platforms as for example Flickr[34]. See Section 3.3.2 for an implementation of a tag cloud in a faceted search scenario.

In case, no metadata is available[35], or the quality of the available metadata is not sufficient, facets can be derived from the content itself.

Deriving Facets from content

Due to the described shortcomings of using unstructured metadata to derive facets, it is often feasible to directly compile facets out of the content itself. This is advisable especially in the media domain. In most cases, structured metadata is lacking, especially if abstract facets should be derived that denote quality aspects of the content. As described earlier, using unstructured metadata is especially challenging

[33] A folksonomy aims to establish metadata for digital resources. Usually, the folksonomy is created by the people who use the digital resource. These people tag the resource with free text. These tags can then be shared with other users [Peters, 2009]

[34] `http://www.flickr.com`, last accessed on Feb 2011

[35] Note that at least some metadata should always exist - for instance the publishing time

due to the heterogenous and subjective nature of the tagging vocabulary. **Within this thesis, facets are derived from content only.** Chapter 5 tackles the derivation of facets from content in detail.

2.3.3 Summary

In this section, the foundations of Information Retrieval (IR) have been outlined. The most common IR models have been discussed and the general IR process has been described. Several evaluation measures have been introduced and the most commonly used model in IR, the Vector Space Model, has been described in detail. Also, in this section, the principles of Faceted Search, a sub-discipline of IR, have been discussed. Special attention has been paid to faceted classification, a technique that is imperative for a faceted search system. Several techniques to extract facets have been discussed while the most important approach for this thesis is to derive facets directly from content. In the next section, the foundations of machine learning and particularly supervised classification are outlined. Generally, facets can be derived from content using either Information Extraction (see Section 2.2.2) or machine learning. The foundations of machine learning are outlined in Section 2.4.

2.4 Machine Learning from Text

Machine Learning aims to derive algorithms that are capable of learning from available data [Langley, 1995]. This naturally covers a broad range of learning tasks; from navigating self-driving cars like recently advertised by Google[36], to create search engines that automatically fit to a user's information need [Smyth et al., 2011], to mining news to predict upcoming events [Jatowt et al., 2009], and to predict which blogs might be of interest for a user in the future. Aside from that, Machine Learning can be exploited to enable humans to learn from data. This is called *Data Mining* [Tan et al., 2006].

Generally, Machine Learning addresses two problems, namely *supervised learning*, and *unsupervised learning*. The goal of supervised learning is to predict unseen data whereas in unsupervised learning, unknown patterns are identified in data. In

[36]The Official Google Code Blog, `http://googleblog.blogspot.com/2010/10/what-were-driving-at.html`, last accessed on Feb 2011

the next section, the specific application of supervised learning for text is described, since this has been the major technique for this dissertation research.

2.4.1 Supervised Learning from Text

Supervised learning from text, also referred to as *text classification*, is a core technique of machine learning that aims to assign documents to a set of pre-defined categories. As the name implies, in supervised learning, a set of examples is given while the examples have already been assigned to their correct target categories. In other words, text classification is the methodology to label unseen textual resources by learning from a set of labeled examples.

In its simplest form, supervised learning consists of (i) a training phase, and (ii) a test phase. In the training phase, a function is learned from a set of training data while the training data consists of a set of already categorized, or *labeled*, examples. More specifically, a hypothesis H is generated which enables the automatic assignment of unseen examples to predefined categories [Nilsson, 1998]. The generation of training samples is typically costly and often needs experts with both learning as well as domain knowledge. This problem will be tackled in Chapter 5.

Supervised learning methods are also referred to as *inductive learning methods* since they derive general domain knowledge from specific knowledge provided by examples from the domain [Bishop, 2008].

In the field of supervised learning for text, or *text classification*, a training example is typically represented by a vector and a target label. From all training examples, a function is learned whose output is either (i) of discrete or (ii) of continuous nature. In case of a discrete output, the function is called *Classifier* and in case of a continuous output, the function denotes a *Regression*. The learned function can then be used to predict the correct label of unseen examples [Bishop, 2008].

In general, regression analyzes how a dependent and one or more independent variables are related to each other [Bishop, 2008]. Based on this relationship, regression tries to predict the value of the dependent target variable based on the independent input variables [Schroeder et al., 1986]. Regression is often applied to predict and forecast trends, or the behavior of the stock market.

Definition of Text Classification

Text classification can be defined as the task of assigning a value to each pair $\langle d_j, c_i \rangle \in D \times C$ where $D = d_1, \ldots, d_i\}$ is a set of documents and $C = \{c_1, \ldots, c_c\}$ is a set of predefined categories Mathematically spoken, the task is to approximate a target function $\Phi : D \times C \rightarrow \{T, F\}$ where T assigned to $\langle d_j, c_i \rangle$ denotes a decision to assign d_j to the category c_i and F denotes that d_j is not assigned to the category c_i. The target function $\Phi : D \times C \rightarrow \{T, F\}$ is referred to as the *classifier* or *hypothesis*, or *model* [Sebastiani, 2002].

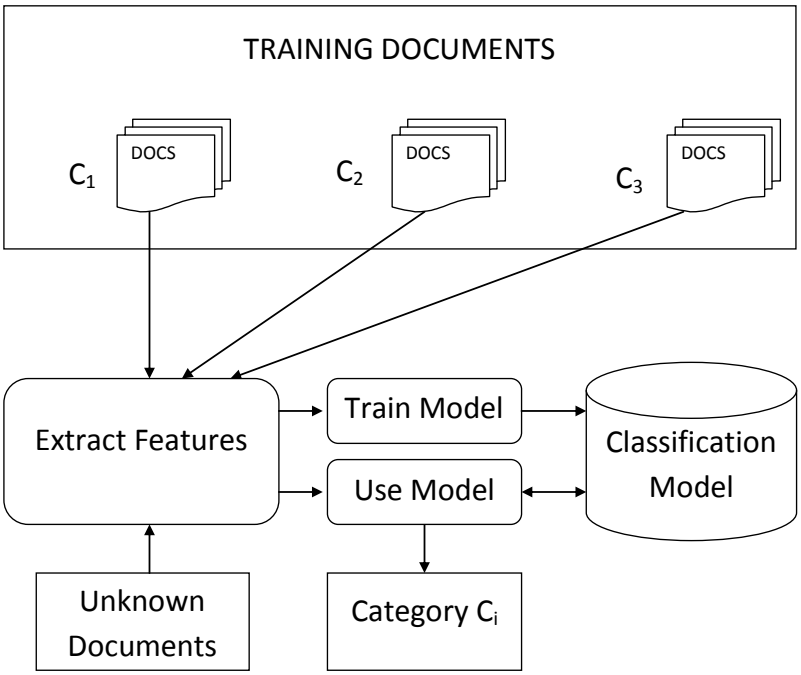
Supervised Text Classification Process

The general procedure for the supervised text classification process is that from the training set, a function, a so-called *model*, is learned. This model captures the characteristic properties of the items of each category. Such characteristic properties are called *features*. These features should exhibit a high discriminative ability so that different categories can be separated from each other [Konchady, 2008]. Each training item in the model exhibits a target label. The trained model is then used to predict unseen items. The result of this prediction is either a continuous value (regression) or a predicted label for an item (classification) [Bishop, 2008]. The general procedure of classification is shown in Figure 2.4 which has been adopted from [Konchady, 2008].

In the classification process, at first, a set of training documents is selected whereas for each category, training documents are needed. For supervised learning algorithms, the availability of a sufficiently large amount of training data is crucial since it is generally understood that this leads to a representative classification model. Also, a balanced number of training examples for each category is feasible in order to prevent favoring one category over the other. If a training set exhibits the same number of training examples for each category, this training set is called *balanced*. However, in practice, training sets are often rather skewed and biased towards one category. See Section 2.4.2 for methods to cope with unbalanced datasets. In document classification, the typically non uniform length of the documents may also be challenging since longer documents may also be favored over shorter texts [Konchady, 2008].

The second step in the classification process is to extract features from the train-

Figure 2.4: The Classification Process



ing documents. These features are then used to train the classification model. Examples for features are tokens, token ngrams, character ngrams, or all types of Part-of-Speech [Konchady, 2008]. The extraction of suitable features is a whole subject of research and it will be shortly outlined in Section 2.4.5.

Based on the extracted features, in the third step, the classification model is trained. Generally, classification models can be divided into two groups: (i) composite models which assign an item to one of n categories, and (ii) multiple models which assign an item to either the positive or the negative category (true or false). Composite models are multi-class classifiers which means that an item has to be assigned to one of n categories based on the likelihood that the item actually belongs to a category. Multiple models are binary classifiers which are used to compute the probability that an item belongs to a category or not [Konchady, 2008].

Finally, a classifier learned on the characteristic features of the categories should be able to correctly predict unseen items. The output of a classifier is either a numeric value (0 or 1) or a probability distribution. In case of a numeric value, the classifier outputs whether the document belongs to a category or not. In case of a probabilistic output, a ranking categorization can be derived that ranks the categories

C according to their probability that they belong to a class c_i [Sebastiani, 2002]. The most likely label for an example is defined by the category with the highest probability (Maximum a Posteriori, MAP). A probability distribution for a sample d_j is given by $p^j = (p^j_1, .., p^j_c), \quad \sum_{k=1}^{c} p^j_k = 1$ whereas c corresponds to the number of classes [Manning et al., 2008].

Note that the output of some classifiers can be mapped to a probability distribution even if they algorithmically output just numeric values (for details on how to map classifier outputs to a probability distribution, see [Wu et al., 2004]).

Types of Text Classification Problems

Text classification problems can be either *single-label* or *multi-label*. Single-label means that a document is assigned to only one category while the categories are mutually exclusive and non overlapping. If a document has to be assigned to more than one category, the text classification problem is multi-label. One also distinguishes between *binary* and *multi-class* classification problems. Binary classification means that only two categories are available: the *positive class* and the *negative class*. More specifically, binary classification models aim to identify whether an example belongs to a class or not. If it belongs to the class, it is assigned to the *positive class* and if not, it is classified into the *negative class*. Naturally, for binary classification problems, training samples for both the positive as well as the negative class are needed. In contrast to binary problems, in multi-class classification problems, c categories are available. In other words, in multi-class classification, the goal is to distinguish between multiple classes. Clearly, binary classification is more general than multi-class classification since multi-class classification with the categories C can be transformed into a number of c binary classification problems under the categories $\{c_i, \bar{c}_i\}$ for $i = 1, \ldots, c$ which are solved independently [Joachims, 2002].

A special category of supervised classification problems is *one-class classification*. In one-class classification, in the training phase, examples are assigned to only one target class. In the classification phase, the unlabeled test items are then assigned to only this one class. The difference between one-class classification and binary classification is that no negative examples are needed for one-class classification. See [Koppel and Schler, 2004] for an example of using one-class classification for authorship attribution.

Applications for Supervised Text Classification

Text classification has been used in multiple different applications. This section provides a listing of text classification applications. While this listing is clearly not exhaustive, it contains some very prominent applications. For a detailed surveys on text classification applications, see [Sebastiani, 2002, Duda et al., 2001].

Organizing large document collections The organization of large document collections is of great importance due to the huge amounts of documents available on the Web. In this context, supervised text classification can be exploited to facilitate the navigability and retrieval of documents on the Web. For instance, blogs dealing with news related events may be categorized according to their main topic. For instance, a political blog may be classified into a category *politics*. Or a blog introducing the newest mobile gadgets like for instance the famous blog Gizmodo[37] may be classified into a category *technology*. Similar applications are the organization of patents into predefined categories [Iwayama et al., 2005] in order to facilitate search[38]

Information Filtering Information filtering is the task of provide people with pieces of information in which they are actually interested in. These interests may be saved in individual or group preferences, so called *profiles*. Information filtering is typically needed on unstructured or semi-structured data[39] [Belkin and Croft, 1992]. A common example for information filtering is the automatic filtering of E-Mails into *spam* or *junk*.

Spam Detection Especially on the Web, the task of spam detection is crucial and therefore it has been tackled by many researchers. For example applications in this context, see [Araujo and Martinez-Romo, 2010, Abernethy et al., 2008]

Tagging Tagging is the task of annotating tokens e.g. with their Part-of-Speech (POS) or their entity type (see Section 2.2.2). For examples of POS tagging using

[37] `http://gizmodo.com/`

[38] Patent retrieval is a whole subject of research which is supported by the NTCIR (NII Test Collection for IR Systems) Project `http://research.nii.ac.jp/ntcir/index-en.html`

[39] An example for semi-structured data are E-Mails since they exhibit both structured headers as well as unstructured textual content

text classification refer to [Spoustová et al., 2009]. Tagging can be tackled as a text classification task whereas the POS or entity type of interest are the target labels and already annotated words form the training set.

Author Identification Authorship identification is the task of determining the author of a text. For this, text classification algorithms learn patterns from the writing style of the author. Based on the learned patterns, it is predicted who actually wrote the text. The main challenge here is to define features capturing the writing style of authors. For examples of author identification using text classification, see [Coyotl-Morales et al., 2006, Koppel and Schler, 2004].

Language Identification Language Identification aims to identify the type of language a given text is written in. Language identification has become increasingly important since on the Web, more and more documents are available in different languages. Language identification is needed for simple applications as for example spell checking, or stemming (see Section 2.2.2) as well as for complex applications like multilingual document retrieval. Language identification can be tackled as classification problem where a classifier is trained on texts in all languages of interest to predict the language of an unseen text. For examples, see [Cavnar and Trenkle, 1994, Dunning, 1994].

Word Sense Disambiguation The goal of Word Sense Disambiguation (WSD) is to assess the sense of an ambiguous word in its context. The context can be a document, a paragraph or a sentence. For instance, the word *bank* has different meanings in a financial context than in a document about a river. WSD can be tackled as a supervised text classification task whereas the contexts in which a particular word occurs serve as training data and the word's sense is the category [Navigli, 2009]. For an example of a WSD approach that is based on text classification, see [Fan and Friedman, 2008].

2.4.2 Text Classification Algorithms

A large amount of different supervised text classification algorithms is available with diverse principles of operation. For instance, neural network based algorithms aim to imitate functions of the human brain (see [Zhang, 2000] for a survey of neural

networks algorithms). Genetic algorithms exploit ideas from the field of evolution; see [Pietramala et al., 2008] for an example of using genetic algorithms for text classification. A large number of algorithms based on probabilities and statistics (see [Kim et al., 2006] for an example). For a review of supervised classification algorithms, see [Duda et al., 2001, Kotsiantis, 2007].

2.4.3 Lessons Learned

Each text classification algorithm exhibits strengths as well as weaknesses. Consequently, choosing the right algorithm for an application task is challenging. As emphasized in the *No free lunch theorem* [Wolpert and Macready, 1997], there is no single classification algorithm that is capable of solving all Machine Learning problems.

When the application task is to solve a learning problem on textual data, it is advisable to use statistic and probability based algorithms. Such algorithms usually achieve the best performance on textual data. In this context, the most common and applicable algorithms are Support Vector Machines, Naive Bayes, K-Nearest Neighbor and Decision Trees.

In this thesis, several supervised machine learning algorithms have been chosen and applied to the problem setting of content facet classification. The selection of the algorithm has been based on the considerations of performance and accuracy. The applied classification algorithms are described in the next sections.

k Nearest neighbor Algorithm

The k Nearest neighbor (k-NN) algorithm [Aha et al., 1991] is a text classification algorithm that takes into consideration the k nearest neighbors of an item in the vector space. An item is classified based on the majority vote of its k nearest neighbors from the training set.

The training phase of the k-NN is rather simple; the algorithm stores the training feature vectors and their class labels. At classification stage, the unlabeled examples, the test vectors respectively, are assigned to the class labels that is most frequent within the k nearest training examples.

This directly leads to the two main drawback of the k-NN; (i) the k-NN algorithm strongly depend on the quality of the used data set and (ii) the parameter k has

a strong influence on the performance of the algorithm. A general rule is that the more noise in the data set, the larger k should be [Witten and Frank, 2005].

Many methods have been proposed to determine the best performing k for a classification task: for instance, in [Arlot and Celisse, 2004], k is chosen based on heuristics and cross-validation. A common approach is to apply bootstrapping. Bootstrapping enables to reduce the classification error by creating new classifiers on a subset of the training data and then estimating the distribution on these subsets [Steele, 2009].

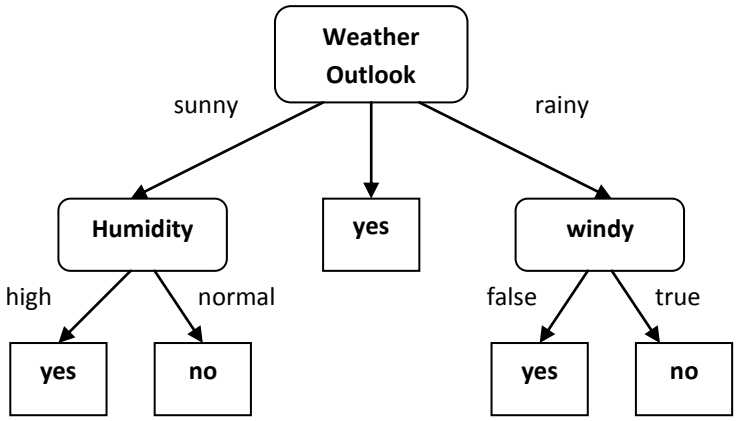
To determine the distance between a test vector and its nearest neighbors, several distance measures have been proposed. The most prominent are (i) the Euclidean distance, and (ii) the cosine similarity (see Section 2.2.3, Equation 2.2). The Euclidean distance $\overline{\mathbf{a_1^{(1)}}\mathbf{a_1^{(2)}}}$ is applicable in dense feature spaces; in text classification however, mostly the feature spaces are highly sparse. In sparse feature spaces, the cosine similarity should be preferred. The Euclidean distance between two vectors $a_1^{(1)}, a_2^{(1)}, \cdots, a_k^{(1)}$ where k denotes the number of features and $a_1^{(2)}, a_2^{(2)}, \cdots, a_k^{(2)}$ is given in Equation 2.9 [Witten and Frank, 2005].

$$\overline{\mathbf{a_k^{(1)}}, \mathbf{a_k^{(2)}}} = \sqrt{(a_1^{(1)} - a_1^{(2)})^2 + (a_2^{(1)} - a_2^{(2)})^2 + \cdots + (a_k^{(1)} - a_k^{(2)})^2} \tag{2.9}$$

Decision Trees

Decision Trees were amongst the first text classification algorithms. Decision trees are tree-shaped structures that represent a set of classifier decisions. From these decisions, rules for classifying a dataset are derived. Each node in the decision tree evaluates an attribute in the data and determines the path it should follow. An example of a decision tree with depth 3 is given in Figure 2.5 that has been adapted from [Witten and Frank, 2005]. There are different approaches to construct a decision tree. Most decision tree algorithms use either a *top-down* approach or a *divide-and-conquer* strategy. In the divide-and-conquer approach, the creation of the decision tree starts at a root node which denotes the so-called *parent* of each other node. Then, for each possible value, a branch is constructed in the tree. This process is repeated for each branch whereas only these instance are used that reach the branch. If all instances at a node have been assigned to the same class, this part of tree is finished [Witten and Frank, 2005]. As split criterion, often a greedy strategy is used. This split criterion can be evaluated e.g. based on the

Figure 2.5: Example of a Decision Tree



mis-classification error or based on the entropy [Rokach and Maimon, 2005].

In the training phase, the decision tree is created on the training data. When a new example has to be classified, the new sample is navigated along the decision tree. A manifold of decision tree implementations exist. See [Safavian and Landgrebe, 1991] for a survey on decision trees. The most popular are the J.48 decision tree [Quinlan, 1993] and the C.45 decision tree. Within this thesis, these two decision tree implementations have been used.

Support Vector Machines

Support Vector Machines that have been introduced by Vapnik in [Vapnik, 1995]. Support Vector Machines belong to the group of linear classifiers. Linear classifiers aim to learn a plane or hyperplane, respectively in order to separate two classes. In a two-dimensional space, the classifier hypothesis is a plane whereas in high-dimensional space, the plane becomes a hyperplane [Granitzer, 2006]. Support Vector Machines are linear large margin classifiers that can be used for classification as well as regression tasks. In the field of text classification, SVMs are among the best performing algorithms [Joachims, 1998].

As described by Vapnik, a Support Vector Machine (SVM) maps input vectors into a high-dimensional feature space using an a priori defined nonlinear mapping. In other words, the input vector space is transformed into a new high-dimensional vector space. In this new space, a linear model can be constructed that actually corresponds to a decision boundary in the original space that is non-linear. This

linear model is called *hyperplane* [Witten and Frank, 2005].

The hyperplane aims at optimally separating the two classes. For this, the hyperplane with the maximum margin between the two classes has to be identified. This so-called *maximum margin hyperplane* separates the classes at maximum. The vectors that have minimum distance to the maximum margin hyperplane are referred to as *support vectors* whereas for each class, at least one support vector exists. The identification of the support vectors and the determination of the parameters is a quadratic optimization problem. This optimization problem is called *constrained quadratic optimization* [Witten and Frank, 2005].

In the following, the mathematical foundations of a linear SVM are outlined, following the tutorial on SVMs by [Burges, 1998]. Given the training data $\mathcal{D}$ consisting of a set of n data points:

$$\mathcal{D} = \{(\mathbf{x}_i, y_i) \mid \mathbf{x}_i \in \mathbf{R}^p,\, y_i \in \{-1, 1\}\}_{i=1}^{n} \tag{2.10}$$

The mathematical formulation of the maximum margin hyperplane is given in Equation 2.11.

$$(\mathbf{w} \cdot \mathbf{x}) - b = 0 \tag{2.11}$$

Note that $\mathbf{x}$ denotes a training example vector, and $\mathbf{w}$ is a normal vector on the hyperplane; $\cdot$ denotes the dot product. b is a numeric parameter that has the be determined by the learning algorithm. The parameter $\frac{b}{|\mathbf{w}|}$ determines the offset of the hyperplane from the origin along the normal vector $\mathbf{w}$. Since an SVM aims to derive the maximum margin hyperplane, the vector $\mathbf{w}$ and b have be maximized. For this, two parallel hyperplanes are created whereas their distance defines the maximum margin. Consequently, this distance should be as large as possible while no training samples should lie between the two hyperplanes. The mathematical formulation for these two hyperplanes is given in Equation 2.13 [Burges, 1998].

$$(\mathbf{w} \cdot \mathbf{x_i}) - b \geq +1, \quad \text{for } y_i = +1 \tag{2.12}$$

$$(\mathbf{w} \cdot \mathbf{x_i}) - b \leq -1, \quad \text{for } y_i = -1 \tag{2.13}$$

The equations for the two hyperplane can be combined resulting in a set of

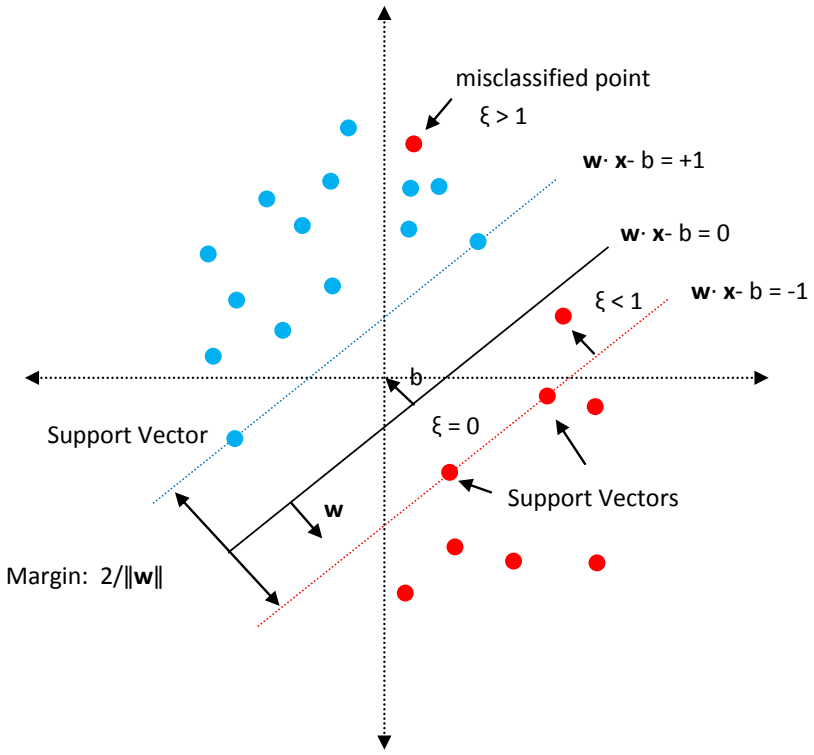
inequalities given in Equation 2.14.

$$y_i(\mathbf{x_i} \cdot \mathbf{w} + b) - 1 \geq 0 \qquad \text{for all } 1 \leq i \leq n \tag{2.14}$$

The points that satisfy the equality 2.13 lie on the hyperplane $H_i = \mathbf{x_i} \cdot \mathbf{w} + b = +1$ with a normal vector $\mathbf{w}$ and perpendicular distance from the origin $\frac{|1-b|}{||\mathbf{w}||}$. The same holds for points that satisfy equality 2.13 - in that case, they lie on the hyperplane $H_i = \mathbf{x_i} \cdot \mathbf{w} + b = -1$ with a normal vector $\mathbf{w}$ and perpendicular distance from the origin $\frac{|-1-b|}{||\mathbf{w}||}$. Consequently, the distance between both hyperplanes is $\frac{2}{||\mathbf{w}||}$. This distance naturally defines the *margin*.

These mathematical connections are shown in Figure 2.6 which has been adopted from [Microsoft Research, 2011]. In practice, there might be no hyperplane that

Figure 2.6: Principles of a Support Vector Machine



separates all of the positive and negative examples. This issue has been addressed by Cortes and Vapnik in [Cortes and Vapnik, 1995]. In this work, they proposed modified maximum margin approach that incorporates examples that have been mislabeled. Their proposed so-called *soft margin* approach derives a hyperplane

that is able to correctly split most of the examples. In other words, the hyperplane maximizes the distance to the nearest correctly placed examples. Additionally, the approach introduces so-called *slack variables*, ξ, that how strong an example x_i is misclassified. The soft margin approach results in a modified formula for the hyperplane which is given in Equation 2.15 [Cortes and Vapnik, 1995].

$$y_i(\mathbf{w} \cdot \mathbf{x_i} - b) \geq 1 - \xi_i \quad 1 \leq i \leq n \tag{2.15}$$

From Equation 2.15 it can be derived that non-zero values for the slack values ξ are penalized by a cost parameter C. The earlier described optimization problem of the SVM then is modified to find a trade-off between a large margin and a small error penalty [Cortes and Vapnik, 1995].

Since data in the vector space is not always linear separable, the so-called *kernel trick* has been introduced that enables to project data into a higher dimensional space. In this case, the dot product is replaced by a (non-)linear kernel function. There are four basic kernels: linear, polynomial, Radial Basis function (RBF), and sigmoid. See [Hsu et al., 2003] for details on these kernel functions. Generally, the performance of a Support Vector Machine strongly depends on the selection of the kernel function and the soft margin parameter C. For text classification, often a linear kernel serves best [Joachims, 1998].

Note that while SVMs have originally been designed as binary classifiers, however, meanwhile also multi-class SVMs exist. See [Duan and Keerthi, 2005] for a review of multi-class methods for SVMs.

Class Feature Centroid Algorithm

The Class Feature Centroid (CFC) classifier [Guan et al., 2009] belongs to the group of centroid based classification algorithms. In centroid based classifiers, classes are typically represented by their centroids. The performance of centroid-based classifiers is good in terms of accuracy and time complexity [Han and Karypis, 2000a]. For the CFC algorithm, a novel centroid weight representation has been introduced that considers both the inter-class term distribution as well as the inner-class term

distribution. The formula for the weight representation is given in Equation 2.16.

$$w_{ij} = \frac{b^{DF_{t_{ij}}}}{|C_j|} \times log(\frac{|C|}{CFt_i}) \tag{2.16}$$

Note that i represents a term, j denotes a class and w corresponds to the term weight of term i of class j. The value b is a constant larger than 1. DFj_{t_i} represents the terms document frequency in a class, $|C_j|$ denotes the number of documents in class j and $|C|$ the number of documents of all classes. The term CF_{t_i} represents the number of classes containing term i. The first component of following equation represents the inner-class term index and the second the inter-term index [Guan et al., 2009].

With this weighting schema, highly discriminant centroids (one for each class) can be derived because terms that occur over all classes are given a low weight and terms that occur in only one class a high weight. This guarantees that the most discriminant terms are assigned a high weight. To compute the similarity of a test document with a class centroid, a so-called denormalized cosine similarity is applied. The similarity between a document vector and a centroid vector is computed using a standard cosine similarity, which is a common similarity measure when dealing with text. However, the centroids are not normalized in order to preserve their discriminant abilities. In [Guan et al., 2009], Guan et al. compare the performance of the algorithm with other centroid-based approaches and with variants of SVMs. All experiments are carried out on the 20-newsgroup and the Reuters-21578 corpus. They reveal that the CFC algorithm outperforms the centroid-based approaches as well as the SVMs. Note that, as outlined in [Sebastiani, 2002], Support Vector Machines belong to the best performing text classification algorithms.

Boosting Algorithms

Boosting algorithms have been introduced in [Kearns, 1988]. Like the earlier described Support Vector Machine algorithm, Boosting is also a large margin classifier. Boosting algorithms combine the decisions of a set of weak learners to predict unseen examples. Weak learners typically only cover parts of the training set; e.g. a weak learner learns only one feature. In Boosting, the strongest of the set of weak learners is assigned the highest weight. See [Schapire, 1990, Freund, 1995] for a detailed explanation of boosting. There exist many variants of Boosting

algorithms; their main difference is how they weight training data items as well as the weak learners' classification hypotheses. A very popular Boosting algorithm is AdaBoost [Freund and Schapire, 1996]. The advantage of Boosting in general is that such algorithms can be easily adapted to different problem settings. See [Granitzer, 2004] for an adaption of a Boosting algorithm.

Ensemble Classifiers

The idea behind ensemble classifiers is that they do not learn a single classifier but they learn a set of classifiers and combine the predictions of all classifiers. The advantage of this methodology is that it should enhance the overall classification accuracy since the decision of an ensemble less depends on the properties of a single classifiers. For this, an ensemble of diverse classifiers is created whereas each classifier is trained on the same training set [Chawla et al., 2001].

2.4.4 Measures for Evaluating Classification Methods

For a meaningful evaluation of classification algorithms, four data sets should be used: a training set, a test set, a validation set, and an evaluation set. A classifier is learned on a training set that contains the labeled items and applied to the test set.

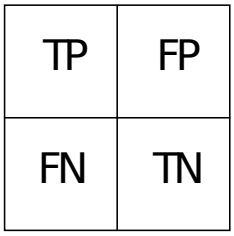
Some supervised learning algorithms require parameters that need to be set beforehand, e.g. before k-NN can be used, the parameter k needs to be set. Such parameters mostly depend on the used data set. Therefore, a specific data set, the *validation set*, can be used to optimize these parameters. Another option is to simply use cross-validation. Cross-validation is a technique to derive the accuracy of a classifier by dividing the data set into n mutually exclusive subsets, so-called *folds*. The classifier is then trained and tested i times whereas training is performed on the data set minus the f_i current fold and evaluated on the f_i fold. Note that i corresponds to the iteration and $i = 1 - N$, N denotes the number of iterations. The advantage of cross-validation is that it can be used to avoid overfitting to a subset of the data [Manning et al., 2008]

The purpose of the evaluation set is to measure the quality of the classifier by common statistical means such as accuracy, precision, and recall. For a detailed overview of such evaluation measures refer to [Sokolova et al., 2006]. In order to get

a reliable prediction of the classifier's performance on unseen data, items from the evaluation set should not be contained in the training set.

A classifier's predictions can be entered in the so-called *Confusion Matrix*. The Confusion Matrix [Kohavi and Provost, 1998] is of size $c \times c$ whereas c corresponds to the number of different categories. Figure 2.7 shows the confusion matrix for a binary classifier (two categories).

Figure 2.7: The Confusion Matrix



From a 2×2 confusion matrix, several performance measures can be derived which are explained briefly in the next sections.

Accuracy

The performance measures accuracy is defined as the rate of correct predictions to all data samples. Accuracy is also sometimes referred to as coverage. The formula for accuracy is given in equation 2.17.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{2.17}$$

True Positive Rate (Recall, Sensitivity)

The True Positive Rate (TPR) encodes the number of samples which were correctly assigned to the positive class in relation to all samples which actually belong to the positive class. Therefore, the TPR corresponds to the estimated conditional probability. The TPR is given in Equation 2.18.

$$TruePositiveRate = \frac{TP}{TP + FN} \tag{2.18}$$

True Negative Rate (Specificity)

The True Negative Rate (TNR) encodes the number of samples which were correctly assigned to the negative class in relation to all actually negative samples. The TNR corresponds to the estimated conditional probability and it is given in Equation 2.19.

$$TrueNegativeRate = \frac{TN}{TN + FP} \tag{2.19}$$

Precision

The Precision is defined as the number of true positives divided by all positive results which consist of both true positives and false positives. The Precision is given in Equation 2.20.

$$Precision = \frac{TP}{TP + FP} \tag{2.20}$$

False Positive Rate (Fallout)

The False Positive Rate (FPR) reveals the number of samples which were wrongly assigned to the positive class even though they belong to the negative class. The FPR is given in Equation 2.21.

$$Precision = \frac{FP}{TN + FP} \tag{2.21}$$

2.4.5 Feature Selection for Classification

Joachims et al. claim in [Joachims, 1998] that classifiers should learn from many features in order to avoid potential loss of information. This is proven by experiments showing that a classifier trained on the worst features usually performs better than random. The availability of many features typically results in a dense feature space. In the field of text classification, document vectors are yet mostly highly sparse which means that the vector contains only a few non zero features.

There exist some algorithm to evaluate the importance of a feature. Prominent examples are the *Linear Correlation (LC)* and the *Mutual Information (MI)*. The Linear Correlation examines if two random variables x_i and y are associated or not. The LC is computed as follows [Guyon and Elisseeff, 2003]:

$$LC(i) = \frac{\sum_{k=1}^{m}(x_{k,i} - \overline{x_i})(y_k - \overline{y})}{\sqrt{\sum_{k=1}^{m}(x_{k,i} - \overline{x_i})^2 \sum_{k=1}^{m}(y_k - \overline{y})^2}} \tag{2.22}$$

where x_i corresponds to the input vector that has the dimension m, and y corresponds to the m dimensional vector that contains the target values.

The Mutual Information is a measure for arbitrary dependency between random variables and has been used extensively in the literature for feature selection [Blum and Langley, 1997, Guyon and Elisseeff, 2003]. In case of this dissertation, this means that the correlation between an input vector x_i and a target vector y is computed. Equation 2.23 gives the Mutual Information for discrete variables [Guyon and Elisseeff, 2003]:

$$MI(i) = \sum_{x_i} \sum_{y} P(X = x_i, Y = y) \log \frac{P(X = x_i, Y = y)}{P(X = x_i)P(Y = y)} \tag{2.23}$$

where $P(X = x_i, Y = y)$ is the joint probability of x_i and y.

2.5 Summary

In this chapter, the theoretical foundations of this dissertation thesis have been outlined. The foundations of information retrieval and faceted search have been discussed and the research connection between faceted search and faceted classification and the proposed dissertation research has been created. This chapter also introduced the concepts behind supervised classification and it described the classification algorithms which have been used for this dissertation research.

Chapter 3

News Retrieval Framework

"I fear three newspapers more than a hundred thousand bayonets"
(Napoleon)

3.1 Introduction

Every day, new potentially useful content is created and published in the media domain. A recent study [Pew Research Center, 2011] published by the Pew Research Center[1] in January 2011 revealed that more and more people declare the internet as being their main news source rather than traditional newspapers. Among the 18-29 years old, the internet has outperformed even television as major source for news. Especially since the advent of Web 2.0, for media consumers, the navigation of the steadily growing amount of user generated media and the discovery of information that best fitting one's personal information need is challenging. Media content is nowadays provided not only by traditional media agencies but also over a variety of social channels like blogs or microblogging platforms - many times far in advance of any traditional news coverage [Granitzer et al., 2010]. As a consequence, many users search for news related events in the blogosphere. A study of blog query logs revealed that 20% of the most popular search queries has been related to breaking news which underpins that people use the blogosphere to gather opinions, thoughts, and discussions about current events [Mishne and de Rijke, 2006].

This development has increased attention towards information quality and cred-

[1] `http://pewresearch.org/`

ibility [Granitzer et al., 2010]. Since the definition of quality and credibility is a rather user specific issue and depends on the actual information need, the judgement whether to trust a resource should lie in the hand of the user. As derived from Chapter 2, the information need of users can be addressed by providing facets in addition to the content itself.

This challenge has been addressed within this dissertation research. For this thesis, **a Web based application framework has been developed that provides content facets in addition to media content**: news repositories as well as social media. This chapter describes the application framework as well as the facets that have been implemented within the framework. Note that the content of this chapter has been published in [Lex et al., 2008, Kienreich et al., 2008].

3.2 The APALabs Framework

The APA Labs framework is an experimental web-based platform fostering the analysis and retrieval of news articles[2] and blogs [Lex et al., 2008]. The goal of APA Labs is to invite arbitrary users to participate in developing, testing and evaluating novel ways to access news agency repositories. APA Labs exploits popular Web 2.0 concepts, for instance the perpetual beta paradigm where the goal is not to create perfect software but to release new feature very fast and to incorporate users at the early stage of the development stage [O'Reilly, 2005]. In APA Labs, users are invited to judge the usability of the visualizations and to provide early feedback to the proposed modules in order to increase user acceptance.

Generally, the APA Labs framework implements a set of visually supported faceted search modules where users can filter media content by geographic location, person names, named entities, and date. The modules are described in Section 3.3.

APA Labs has been implemented in Java as a web application based on J2EE technology[3]. A client-side rich internet application exploits JavaScript and AJAX technology to communicate with an Apache Tomcat Web Server Version 5.5[4] pro-

[2]The archive of the Austrian Press Agency (APA) contains around 50 million news articles in German language and grows by approximately 10.000 articles per day.

[3]Java 2 Platform Enterprise Edition Specification, v1.4, `http://java.sun.com/j2ee/j2ee-1_4-fr-spec.pdf`, last accessed Jan 2011.

[4]The Apache Software Foundation. Apache Tomcat, `http://tomcat.apache.org`, last accessed Jan 2011.

viding content through Java Servlets and Java Server Pages.

The APA Labs server features a central request handler servlet that accepts and forwards requests made by registered sessions. A session is registered on creation and assigned light-weight state data as for example the number of search results by means of session attributes. Due to performance reasons, heavy-weight, persistent state data as for example a news article search result set is stored in a separate repository that can be accessed by the particular session.

In Figure 3.1, the system architecture of APA Labs is shown. The functional components of APA Labs are described in more detail in the next sections.

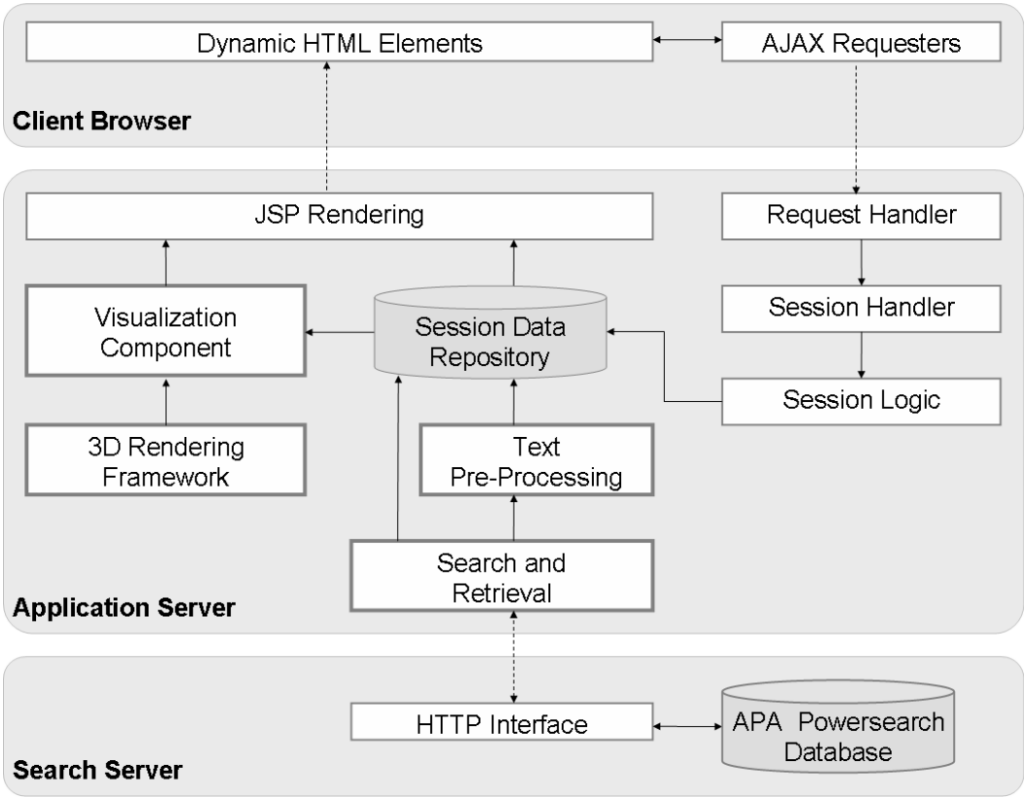


Figure 3.1: APA Labs: System Architecture

3.2.1 Search and Retrieval

The search and retrieval component is responsible for accessing and searching the news article repository of APA. This is accomplished by querying the APA PowerSearch engine [APA-IT, 2011], a generic search engine maintained by APA. The

APA PowerSearch engine supports Boolean queries over a HTTP-based query interface and returns a list of news articles that are ranked by relevance. For each news article, a set of metadata is provided, namely title, publishing medium, and publication date. The search and retrieval component of APA Labs stores the resulting news articles and the metadata in the session-specific data repository. Due to performance reasons on the client side of the application, news article details as for example the article content are loaded only when necessary. For instance, the article content is only required in a full text view but not for displaying the search results in a result list.

3.2.2 Keyterm Extraction and Named Entity Extraction

The keyterm extraction and Named Entity extraction component identifies key terms as well as Named Entities. For this, the content of news articles is first pre-processed so that relevant noun phrases are detected. This pre-processing step exploits stop word lists, stemming, and language and domain specific heuristics. Second, from the resulting noun phrases Named Entities are extracted based on term statistics, domain specific heuristics, and gazetteer lists. Currently, keyterm extraction and Named Entity extraction component is able to identify persons, geographic locations, Web addresses, date and time expressions, and also pre-defined topics of interest.

3.2.3 Rendering Framework

The server-side generation of the visualizations uses the Java bindings for OpenGL (Jogl) [Segal and Akeley, 2006], a rendering framework. The rendering framework enables to generate complex visualizations very efficiently. The computed visualizations are then sent to clients in form of a combination of compressed images and structured image maps identifying mouse interaction areas.

3.2.4 Visualization

The visualization component is an abstract container that enables to create graphical user interfaces to visually provide and analyze search results. In a concrete implementation of a visualization component, it is defined which types of extracted

entities are used for the analysis and how they are visually represented. As depicted in Figure 3.1, the required data is then loaded from the session data repository that holds the results of the search and retrieval component as well as the results of the pre-processing component. The Jogl based visualization component finally generates the resulting visualization and delivers the result back to the client. Due to the modular structure of APA Labs, new visualizations can be easily integrated by simply re-implementing the visualization component.

3.2.5 Feedback Component

The feedback component enables to collect and evaluate user feedback. In general, many Web 2.0 applications gather user feedback to analyze the opinion of the community. Insights gained from user feedback enables the application providers to react on usage trends and user approval. In APA Labs, the feedback component is added to each visualization module. Users have the possibility to rate the particular visualization module in respect to functionality, design and usability on a 5-point Likert scale [Likert, 1932]. Further, users can comment on the visualizations via E-Mail or a feedback form. A visualization module manager at the Austrian Press Agency then collects the user feedback which serves as a basis for further decisions like improving the visualization adding the experimental prototype visualization to the business product range of APA.

3.2.6 Interfaces

All components and modules of APA Labs are connected via well-defined interfaces; the first main interface links the search server with the application server while the second main interface links the framework itself with its visualization modules. Additionally, the framework consists of several components for data exchange and session handling. The search server communicates with the application server over a Representational State Transfer Architecture (REST) [Fielding and Taylor, 2002] interface that is accessed via the HTTP protocol using a clear syntax. Following the interface definitions, new visualizations can be integrated in the framework without knowledge about the underlying logic of the whole application.

3.3 Modules

The APA Labs[5] framework consists of several custom modules that enable to navigate and filter search result sets. Part of the functionality of APA Labs is rather conventional: users can search for news articles as well as blogs, they can navigate search result sets through relevance ranked lists, and they can access the fulltext. Additionally, the user interface of APA Labs provides a set of experimental visualizations with consistent design and user interaction. These visualizations share the ability to provide an alternative way to formulate an information need, to navigate a result set or to analyze article content.

The general visual layout of the platform is illustrated in Figure 3.2.

Figure 3.2: APA Labs: Overview

There are two different types of visualization modules: (i) modules that operate on a set of documents and (ii) modules that analyze a single document. The modules are implemented as classes within the application framework and, like described

[5]APA Labs, `http://www.apa.at/labs`, last accessed Feb 2011

earlier, access search result sets, extracted Named Entities and rendering components through well defined interfaces. Note that most of the modules available in APA Labs operate on news articles provided by the Austria Press Agency APA.

3.3.1 Geospatial Visualization

Geospatial visualizations display information entities that correspond to geographical locations on appropriate maps [Scharl and Tochtermann, 2007]. Geospatial visualizations are a natural extension for systems analyzing news articles because most news articles reference one or more geographic locations. More than 85% of all articles available in the archives of the Austrian Press Agency contain at least one geographical reference. The geospatial visualization present in APA Labs extracts geographic locations from a set of documents resulting from a preceding search query. Figure 3.3 displays the Geospatial Visualization of a search result set obtained in

Figure 3.3: APA Labs: Geospatial Visualization

the APA Labs framework. Colored three-dimensional cones have been positioned on a map of Austria representing locations mentioned in one or more articles. The size of each cone encodes the number of occurrences identified for its location. The cones are rendered using a semi-transparent material to avoid occlusion effects. Moving the mouse pointer over a cone displays the name of the location and the number of references identified for it in form of a tool tip window. Clicking on a location instantly filters the search result set to contain only articles referencing the selected location. One benefit of the Geospatial Visualization is the ability to identify geographical hot spots for a particular topic at a glance. Another benefit is the ability to quickly refine the search results by region.

3.3.2 Tag Cloud Visualization

Tag clouds are text-based visual representations of a set of words (tags) usually depicting tag importance by font size. The popularity of this type of visualization has steadily grown due to recent trends in social and collaborative software. In contrast to many other types of visualizations, tag clouds do not use real-world models or metaphors. A tag cloud is a visual abstraction and thus suitable for visualizing information entities of arbitrary types. Tag clouds are especially useful for topical browsing [Kuo et al., 2007] and consequently also for browsing of news articles.

Tag clouds have become very popular in Web 2.0 applications, such as del.icio.us and Flickr. For details on the layout algorithm used for the tag cloud implementation in APA Labs, a technical evaluation and a user study of different tag layouts, refer to the work of my colleagues [Seifert et al., 2008]. The tags for the tag could

Figure 3.4: APA Labs: Tag Cloud Visualization

visualization in APA Labs have been derived by extracting Named Entities and important keywords from news article search results. More specifically, the tags denote persons, locations, or general important terms. The font size of a tag corresponds to its importance: tags that occur more often in a result set exhibit a larger font size and they are placed in the middle of the tag cloud. Moving the mouse pointer over a tag displays a tool tip with the tag name. Clicking on a tag instantly filters the search result set to only news articles that contain the particular tag. The main advantage of the tag cloud visualization is that users can easily identify the major topics, persons, and locations of a news search result set. Besides, the user can visually filter the search result set by the most important words; this enables to get

a profound overview of the contents provided within the resulting news article set.

3.3.3 Parliament Visualization

The Parliament Visualization module integrated in the APA Labs framework extracts members of the Austrian parliament from a news article set resulting from a search query in the APA repository. The extracted Member of Parliament are then depicted in a three-dimensional visualization of the actual parliament. This enables users to instantly determine which members of the Austrian government and parliament have been mentioned in the context of a search result set.

The parliament visualization can therefore be used to observe impact statements made by persons of public interest in the media. In the context of APA Labs, these persons of public interest are the leading politicians of Austria.

Figure 3.5 displays the Parliament Visualization of a search result set obtained in APA Labs. The basis of the Parliament visualization is a three-dimensional model

Figure 3.5: APA Labs: Parliament Visualization

of the actual Austrian parliament. Each Member of Parliament has a seat in one of the curved rows of seats in the image. In the foreground, a row of seats is reserved for the government's ministers. A three-dimensional icon in the shape of a human is placed on each set and oriented to face the point of view of the observer. Each icon corresponds to either a Member of Parliament or a minister. For each seat, the name and the political party association of the person holding it is known. Initially, all icons are colored gray and of the same size. For each person mentioned in the search result set, the according icon is colored in the colors of that person's political party. The size of the icon corresponds to the number of occurrences of the particular

politician in the result set. Moving the mouse pointer over an icon displays the name of the Member of Parliament or minister and the number of occurrences in a tool tip window. Clicking an icon instantly filters the search result set to contain only articles mentioning the selected person. The main benefit provided by the Parliament Visualization is the ability to identify which politicians are associated with a specified topic.

3.3.4 Roundtable Visualization

The Round Table Visualization shows the leading person of each political party in Austria. Similar to the earlier described Parliament Visualization, the Round Table Visualization extracts the leading politicians from a search result set. The extracted politicians are then depicted in a three-dimensional visualization of a round table, a common seating arrangement in political discussions. Figure 3.6 gives an example of the Round Table Visualization. The figures in the three-dimensional visualization

Figure 3.6: APA Labs: Roundtable Visualization

are colored according to the particular party affiliation. In front of each figure a label providing the name of the politician is placed. The size of the figures corresponds to the number of occurrences of the particular politician in the search result set. The names of the politicians and the exact number of hits are also available in a tool tip window. Additionally, when moving the mouse pointer over a figure, a small tag cloud is shown that contains the most important names, geographic locations and terms in the context of this politician. These entities are extracted from a subset of the search results whereas this subset only contains that mentioned both the search query term and the particular politician. This subset can also be obtained

by clicking the politician's figure which results in a filtering of the original search result set.

The main advantage of the Round Table Visualization is that it enables users identify who of the top Austrian politicians is associated with a specific subject in the media while this subject is expressed as query. It can therefore serve as tool to analyze the political discourse of Austria's parties. Since for each politician, also the most important keywords are available, users can easily assess which topics are covered by each politician in this context. Naturally, the politicians can be easily changed in the visualization in order to reflect the actual political situation in Austria.

In the course of the development of the APA Labs framework, other sources than German news articles have been included in the analysis. More specifically, a crosslanguage blog mining and blog analysis module has been included in the framework in order to monitor the news response in the blogosphere. As a visual support for this monitoring process, the so called *Blog Trend Visualization* has been proposed [Juffinger and Lex, 2009]. The blog mining and blog analysis module as well as the Blog Trend Visualization will be described in more detail in Section 3.3.5 and Section 3.3.5 since this work has been a first approach towards a deeper analysis of blogs especially in the news media domain.

3.3.5 Blog Mining System

The blog mining system consists of a high performance multilingual blog miner as well of a Blog Trend Visualization. The blog mining system incrementally loads and parse blogs whereas the blogs have to be registered in the system beforehand.

Blog mining exploits techniques from the fields of Web Mining and Text Mining. Additionally, in blog mining, principles from social network analysis are used to assess structure and content of blogs [Juffinger and Lex, 2009].

When blogs are parsed, their unstructured content is transformed into structured blog representations which further are analyzed by text mining components. In case of the blog mining system used for APA Labs, a semi automatic blog parser is used that applies relative XPath queries to extract e.g. single blog posts, or author information directly from the DOM tree [Kowalkiewicz et al., 2006]. This step results in a set of structured blog posts per blog whereas from each post, also

its language, the most important Named Entities, title and content are extracted. Title and content are then indexed in language specific blog indices with the search and indexing framework Apache Lucene[6]. For details on the blog mining system, refer to the work of my colleagues [Juffinger et al., 2009c].

Blog Trend Visualization

The Blog Trend Visualization implemented in the APA Labs framework shows German APA news articles as well as German, English, Spanish, and Italian blog posts over time. To obtain these blog posts, the German search query is translated to English, Spanish, and Italian based on Wikipedia statistics, as outlined by my colleagues in [Juffinger et al., 2009b]. First, the search query is used to search the German Wikipedia index. This results in a set of German blog posts. From the top 50 results, the linked Wikipedia entries for each target language are extracted resulting in a set of relevant Wikipedia entries in each language. From these entries, the most significant terms are extracted which are further used to query the language specific blog index. The resulting blog posts are then displayed in the Blog Trend Visualization in combination with the news articles.

Figure 3.7 gives an example of the visualization for the search query term *Bush*. The time period is limited to the last 60 days. Three days are summarized and summarized by a colored bar. Symbols for blog posts and articles are shown on the bar if search results are available for this timeslot. The blog posts are shown

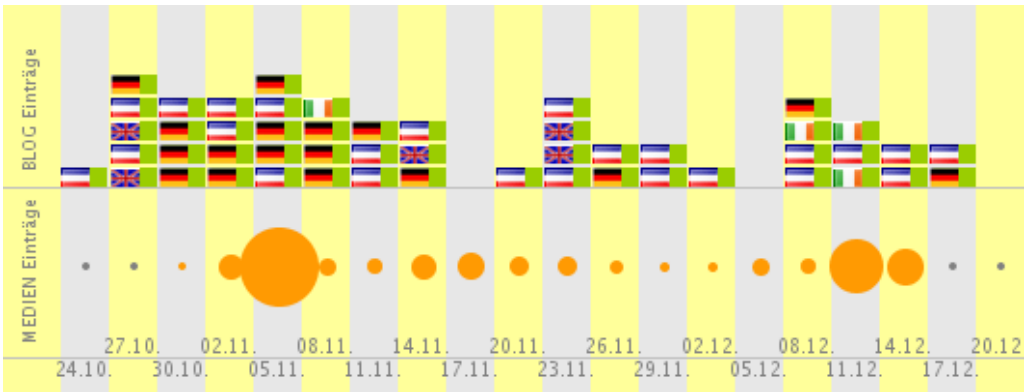


Figure 3.7: APA Labs: Blog Trend Visualization - Visualization for query Bush

in the top section of the Blog Trend Visualization. Each blog post is represented

[6]Apache Lucene, `http://lucene.apache.org`, last accessed Jan 2011

by an icon consisting of a green rectangle and a flag that denoting its language. In the example 3.7, German, English, French and Italian blog entries are available for the query *Bush*. Moving the mouse over a blog post icon shows a tooltip window that contains the title of the post. Clicking a blog post icon opens the link to the external blog site in a separate browser tab.

The APA news articles are represented by orange circles in the bottom of the Blog Trend Visualization. The size of a circle corresponds to the number of articles containing the query term within the particular timeslot. Following the general interaction concept of APA Labs, clicking an orange circle filters the original search result set to news articles that have been published in the selected timeslot.

3.4 Summary

This chapter presented the experimental news and blog retrieval framework APA Labs. The APA Labs framework is designed as a combination of a rich internet application with a modular system of interactive visualization modules exploiting server-side entity extraction and three-dimensional rendering.

The web-based platform combines free text search in new articles from the Austrian Press Agency APA with a visual analysis of the search results. The proposed visualization modules enable users to interactively filter the media content by different content facets, namely geographic location, person names, important keyterms, and date. Aside from news articles, APA Labs also integrates German, English, Spanish, and Italian blogs for a specific module, the Blog Trend Visualization.

The APA Labs platform enables APA to evaluate the user acceptance of new services and to get user feedback early in the development stage. Due to its free online availability [7], APA Labs serves both as generator for public awareness to APA and its products as well as initiator for innovation. Therefore, APA Labs is mutually beneficial for both the company APA and the users of APA Labs.

The development of the APA Labs framework raised the question which further content facets can be derived within the media domain and especially in social media. **This actually is a central question for this thesis.** The following chapters identify, define, and assess such content facets in traditional, Online, and social media.

[7] APA Labs, `http://www.apa.at/labs`, last accessed May 2011

Chapter 4

Content Facet Assessment

> *"The idea is to try to give all the information to help others to judge the value of your contribution; not just the information that leads to judgment in one particular direction or another"* (Richard Feynman)

This chapter outlines the theoretical backgrounds of the assessment of content facets. The most important approaches are described to extract and generate facets from the content of media documents with special attention to social media and blogs. This provides insights related to Research Question 2, *"Into which types of content facets can blogs be categorized?"*, because this chapter derives a categorization of facets into two types, namely facets capturing the topical aspects of a media document and facets that go beyond topical relevance.

The contribution of this part of the thesis are: (i) a literature review of the most important approaches for content facet assessment, (ii) a derivation of a classification of facets into two types of content facets, and (iii) a discussion of cross-domain properties of content facets. From these theoretical considerations, implications for choosing appropriate features and algorithms to assess content facets are derived.

4.1 Definition of Content Facets

What is a Content Facet?

In general, facets are commonly understood as a particular aspect of a digital resource (see Section 2.3). The goal of this thesis is to assess such facets of dig-

ital content with the following restriction: the facets are assessed solely from the resource's content. Formally, this results in the following definition of the term *Content Facet*:

Definition 4.1. A Content Facet F describes a specific aspect of a textual resource R. The content facet F is derived only from the content of R.

If not declared otherwise, this definition of Content Facet is used for this thesis.

> **Example:** Let a set of blogs be the textual resource R. R contains blog posts written by two authors A and B who posted about the topic T. The blogs of A are of low quality Q_{low} whereas the blogs of author B are of high quality Q_{high}. Consequently, these blogs exhibit the content facets F_{author}, F_{topic} and $F_{quality}$. To assess the content facets F_{author}, F_{topic} and $F_{quality}$ only the content of R is used while for example the blogger's social network, or her user profile is not considered.

Content Facets as defined in this dissertation therefore differ from other facets in respect to from which part of information they are assessed.

4.2 Categorization of Content Facets

As Daniel Tunkelang, Principal Data Scientist at LinkedIn[1], described in his blog *The Noisy Channel*[2], that facets can be used for two search tasks:

> *"I like the use of facets in order both for elaborating my initial information need and as guidance for exploration"* [Tunkelang, 2010]

The implication of this statement for this thesis is that content facets can be divided into at least two types: Firstly, content facets can be used to re-formulate and refine keyword based, topical information needs. Secondly, content facets enable users to discover new aspects of content.

The influencing position paper *"What should blog search look like?"* by Marti Hearst et al. [Hearst et al., 2008] also proposed different types of content facets. In

[1] LinkedIn, `http://www.linkedin.com`, last accessed Feb 2011
[2] The Noisy Channel, `http://thenoisychannel.com`, last accessed Feb 2011

this paper, three key tasks in blog search are proposed that should serve as guidelines for future social media search tools. The first key task should enable to find out what people *"think or feel"* about a topic over time. The second task is to *"find useful information"* that was published in blogs in the past and the third task is to *"find good blogs and authors"* to read.

For this thesis, especially the second and the third task are relevant: blog search engines should not only rank blogs by relevance but also by their usefulness and their quality.

This dissertation research therefore aims to provide two types of content facets:

1. **Type 1:** Content Facets to refine and re-formulate keyword based topical information needs
2. **Type 2:** Content Facets to explore aspects of media content that go beyond topical relevance and that denote content quality aspects

4.2.1 Related Work on the Categorization of Content Facets

The proposed distinction of content facets into two types is actually corroborated by a recent tutorial held at the World Wide Web conference in late 2010 [Serdyukov, 2010]. In the tutorial, the task of classifying search results from Enterprise and Desktop Search has been divided into two subtasks: (i) a topical categorization, and (ii) a non-topical categorization. For the non-topical categorization, they proposed the following classification tasks:

- Classification by Genre
- Classification by Reading Difficulty
- Classification by Sentiment
- Build Classification Models: Objective versus Subjective, Positive versus Negative
- Classification by Location: Geotagging, exploiting Location Metadata

The topical classification task aims at grouping search results by their main topic. This enables users to filter the search results by the content facet topic. The non-topical classification tasks cover the content facets genre, sentiment, location, as

well as content quality - indicated by the content's reading difficulty, or e.g. its objectivity.

Summing up, the tutorial underpins the ideas for the content facet assessment in this dissertation research since facets are also divided into two types:

- Topic Oriented Content Facets
- Topic Independent Content Facets

4.2.2 Topic Oriented Content Facets

As the name implies, topic oriented content facets describe the topical aspect of a digital resource. The challenge here is to identify the most relevant topic of a resource since in the media domain, and especially in social media, the topical diversity is large; bloggers e.g. may write about any topic that comes to their minds and the topic landscape is dynamic and steadily changing [Sun et al., 2007].

In the course of this thesis, two topic oriented content facets have been assessed in traditional as well as social media:

1. Topic
2. Genre

Even though in [Serdyukov, 2010], the facet *genre* is regarded as being non-topical, in this work, the facet *genre* is tackled as being a topic oriented facet. Naturally, genre and topic are not the same and some topics may occur over different genres. However, in the media domain, genre and topic are closely related since genres in news often are newspaper categories like sports, politics, etc. This viewpoint is also corroborated by Finn et al. [Finn and Kushmerick, 2006] who found that genre and topic *"in practice [. . .] partially overlap"*.

Topic

The content facet topic can be used to browse and navigate large document collections. Within this thesis, the content facet *topic* is defined as follows:

Definition 4.2. Topic: Let T be the topic of a textual resource R. A topic T is a general subject area as for example Austria or computers [Manning et al., 2008].

The textual resource R is about a topic T if the main statement of R describes or reports about the topic T.

If not otherwise stated, this dissertation refers to this definition of the content facet *topic*.

> **Example:** A news article tackles the political riots in Egypt as T_1. The article describes the political backgrounds of former president Mubarak as T_2 as well as the implications of the riots for the future of the country Egypt T_3. Even though in this example, several topics, namely T_1, T_2, and T_3 are tackled, the main topic $T_{main} = T_1$.

Challenges As indicated in the above example, the challenge in the assessment of the content facet *topic* is to identify the actual main topic T_{main} of a resource R since R possibly tackles more than one topics. This challenge can be addressed via unsupervised Machine Learning or Term Extraction. These techniques are outlined in this context in the next paragraph.

Techniques to assess the Content Facet Topic There exist a number of techniques to assess the main topic T_{main} of a digital resource R. These techniques either (i) make use of prior knowledge about the resource or (ii) they extract topics in an unsupervised way.

In a rather simple form, the unsupervised extraction of topics can be achieved via **Term Extraction**. Term Extraction in this context exploits Information Extraction (IE) methods to extract keywords from a document collection. These keywords further serve as content facets (see Section 2.2.2 for a description of Information Extraction). Through the application of IE, for example the most relevant nouns can be extracted that represent the main topic T_{main}.

Another approach in this context is to combine techniques from Machine Learning, namely, **Clustering and Cluster Labeling**. Generally, Clustering is an unsupervised Machine Learning technique that aims to group together documents. These groups are referred to as *document clusters*. Note that the term *unsupervised* denotes that no a-priori knowledge is available about the collection that is subject to the clustering process.

As described by [Manning et al., 2008], the documents in a cluster should exhibit a high similarity whereas the documents of different clusters should have a low similarity. For a more extensive description of unsupervised learning and clustering refer to e.g. [Manning and Schuetze, 2003]. In case of the topic content facet this translate into grouping of documents that tackle similar topics.

The idea to apply clustering to assess the content facet *topic* is based on work by Cutting et. al. in [Cutting et al., 1992]. In this work, an approach called *Gather/Scatter* has been introduced whose goal is to support users in search tasks. In the Gather/Scatter approach, the search results are first divided into groups (clusters) of documents. Consequently, the document clusters serve as facets in this scenario. The resulting document clusters can then be used to narrow a search result set by selecting an appropriate cluster. See [Hearst et al., 1995] for an implementation of the Gather/Scatter principle in a search system.

One of the challenges in this context is to identify meaningful descriptions of the resulting topic clusters. These descriptions typically indicate the main topic T_{main} of a cluster and serve as topic content facets. This approach is called *Cluster Labeling*. Cluster labeling is a technique to compute meaningful labels for document clusters [Manning et al., 2008]. In cluster labeling, statistical techniques are applied to derive the most important terms of a document collection. The rationale behind that is that the most important terms typically best represent the main topic T_{main} of a cluster. For instance, de Winter et al. [de Winter and de Rijke, 2007] describe a methodology to organize blog posts into a number of facets using clustering. Given a set of blog posts relevant to a topic, several cluster labeling methods for identifying facets to a blog post's topic have been evaluated.

Generally, there exist a broad range of Cluster Labeling techniques. For instance, in [Sabol et al., 2009], clusters are labeled based on a selection of the terms with highest term weight in a cluster centroid.

Also, external knowledge can be beneficial for Cluster Labeling. Dakka et al. [Dakka and Ipeirotis, 2008] proposed a methodology to organize content into hierarchical facets using unsupervised methods. In a study they found that often meaningful facet terms are missing in text documents. As a solution, external resources can be exploited to identify useful facet terms. For instance, Carmel et al. [Carmel et al., 2009] proposed to use external knowledge in form of a thesaurus derived from Wikipedia. The thesaurus is used in a post-processing step to im-

prove the quality of the cluster labels resulting from a statistical labeling procedure. Another example has been proposed in [Muhr et al., 2010], where cluster labeling has been implemented based on hierarchical relationships whereas knowledge about the structural properties of the data is also included. If no external knowledge is available, measures like the *Jensen-Shannon divergence* [Carmel et al., 2006] can be applied to improve the quality of cluster labels. Note that the Jensen-Shannon divergence indicates the difference between two probability distributions [Li et al., 2008]. Therefore, it can be used to measure the statistical difference between document clusters. For more related work on cluster labeling, refer to [Muhr et al., 2010].

Using clustering to extract content facets can be beneficial specifically in the media domain. Typically, the amount of news articles, online news, and social media content is huge and steadily growing. Consequently, it is appealing that new candidate facets can *automatically* be identified from a large set of unseen documents without any prior knowledge about the data.

However, this is also the main disadvantage since the outcome of a clustering procedure is not predictable due to its unsupervised nature. Besides, the identification of meaningful and descriptive cluster labels still remains a difficult problem [Hearst, 2006b].

To overcome these disadvantages, supervised machine learning, namely **topic classification**, can be used. In topic classification, documents are classified into a set of known topics. For instance, media content can automatically be classified into seasonal topics like e.g. *Olympic Games* or *elections*.

Yet in blogs, topic classification is challenging in respect to the following aspects:

- The type of content: Compared to standard objects in text classification, blogs do not consist of only one document but of a set of single blog posts.
- The recency and dynamic nature of content: Blogs are typically frequently updated with new blog posts and therefore highly dynamic compared to standard documents or Web sites.
- The topical diversity: Bloggers may write about any topic which comes to their minds.
- The topic drift: Topics described in e.g. the blogosphere change over time.

The implications of these challenges for this dissertation research are that especially in blogs, there is a need to **use efficiently trainable classification algorithms** since classification models need to be continuously retrained in order to address the topic drift and the dynamic nature. Besides, in blogs, content facets should be assessed on blog post level since each blog post can be regarded as single document that might tackle a different topic.

Genre

The genre of a document is an aspect of the document that differs from the document's topic [Lim et al., 2005a]. Within this dissertation research, the content facet *genre* is defined as follows:

Definition 4.3. Genre: Let G be the genre of a textual resource R. A genre G is a grouping of textual resources based on defined similarities in respect to functional purposes [Chen and Choi, 2008]. The textual resource R can be assigned to G if the content of R belongs to the genre G.

If not otherwise stated, this dissertation refers to this definition of genre.

> **Example:** A news article describes a football match and some celebrities from politics and the entertainment sector who attended the match. This example can be assigned to several commonly agreed upon newspaper genres, e.g. G_1 Sports, G_2 Politics, and G_3 Entertainment.
>
> The earlier introduced content facet *topic* T can be combined with G in retrieval scenarios [Stein et al., 2010]: for instance, a media consumer first searches for a specific topic T and then refines the search result set by the content facet *genre* G.

Challenges The first challenge in the assessment of the topic oriented content facet *genre* G is to avoid that instead of a resource's R genre G, its topic T is assessed. Therefore, to assess G, there is a strong need to be as topic independent as possible in respect to used features and algorithms.

A common approach towards genre assessment is to use supervised machine learning techniques. In this approach, the resource R is classified into a set of predefined genres G_i. Therefore, the second challenge in this context is the availability of meaningful and commonly agreed upon genre categories.

In the media domain, a solution would be to use commonly agreed upon newspaper genres as for example *sports* or *politics* as genres G_i. Another solution in case of blogs is to exploit the genre information provided in so-called blog directories[3]. In general, blog directories categorize blogs into either flat (e.g. BlogFlux[4], accessed on Nov 22, 2010) or hierarchical genre categories G_i (e.g. BlogCatalog[5]). The main advantage of such blog directories is that the genres typically are well defined and commonly agreed upon.

Since the use of supervised Machine Learning for genre classification requires the availability of a sufficiently large amount of training data, this naturally states the third challenge in the context of assessing the content facet *genre* G.

Techniques to assess the Content Facet Genre As mentioned earlier, the assignment of a resource R to the content facet *genre* G can be tackled by classifying documents into a set of predefined genres G_i. For supervised classification, it is crucial to choose characteristic features. In the following, selected approaches towards genre classification are outlined with focus on the used features.

In general, for genre classification, various features can be exploited. Lim et al. describe in [Lim et al., 2005b] five distinct sets of features to automatically classify the genre of web documents. These five feature sets incorporate the information from the URL, HTML tags, token information, lexical information and the document structure.

In respect to the media and blog domain, usually little genre information is encoded in the URL. Also, HTML tags and structural information as for example page impress features, such as the amount of advertising per page, are not applicable to assess the genre G of blogs, because there is no evidence that e.g. the amount of advertisement is different over various genres in blogs [Juffinger et al., 2009a]. Token information and lexical information can definitely be exploited to assess G since some words or characters are more likely to be used in certain genres. For instance, considering the media domain, probably different words and expressions are used between the genres entertainment and economic news.

In [Lee and Myaeng, 2002], genre classification is addressed via a word statistics

[3]Note that a blog directory organizes blogs into predefined categories - like Yahoo! Directory does for Web pages.

[4]`http://dir.blogflux.com`

[5]`http://blogcatalog.com`, accessed on Nov 22, 2010

based approach. Besides, they introduce a special term weighting scheme that consists of a combination of variants of the Document Frequency (DF). This weighting scheme aims to assign terms that indicate the genre of a document a high weight.

Stamatatos et al. [Stamatatos et al., 2000] propose to exploit common word statistics for genre classification. More specifically, word frequencies and punctuation marks are used whereas special attention is paid to derive a language and domain independent methodology.

Finn et al. [Finn and Kushmerick, 2006] define genre as being the style of a document's textual content. They address the genre classification challenge with a machine learning approach that evaluates different feature sets. As features, they suggest bag-of-words, parts-of-speech statistics with stemming and stopword removal, and shallow text statistics as for example the average sentence length. In their experiments, it is also investigated whether genre classifiers can be used to classify documents in different topical domains. They investigate whether genre classifiers can be transformed from one topical domain to another topical domain. For this, a genre classifier is trained on one topical domain and tested on another topical domain. As datasets, they use the movie reviews corpus[6] as well as a restaurant reviews corpus. Their experiments revealed that in a single domain setting, they achieve an average classifier accuracy of 76.8% with bag of words features. In the cross-domain setting which is referred to as *domain transfer* in the publication, they achieve a classifier accuracy of only 47.8%.

This corroborates **two key points of this dissertation research:**

- A document's genre is not completely independent of the document's topic
- In a heterogeneous topic landscape like the blogosphere, there is a strong need to use topic independent features.

4.2.3 Topic Independent Content Facets

Topic independent facets do not depend on the topical content of a document collection. Such facets can be temporal and date related facets, and facets denoting how a document has been written and what major feeling the document covers.

[6]Movie Review Data, `http://www.cs.cornell.edu/people/pabo/movie-review-data/`, last accessed March 2011

These facets can be used as content quality indicator and to assess the opinions and personal feelings of the authors who produced the content. In this thesis, four topic independent content facets have been assessed:

1. Objectivity
2. Emotionality
3. Credibility
4. Quality

This section further describes methodologies to derive topic independent content facets in terms of related work. The selection of the described topic independent content facets has been based on facets that have been actually implemented in the course of this dissertation.

Challenges

The assessment of topic independent content facets is challenging, especially when rather abstract content facets as for example quality have to be assessed. This is due to several reasons:

- The definition of an abstract content facet as for example quality strongly depends on the user's individual definition of quality and on how she uses the information [Eppler, 2003].
- The identification of topic independent features describing the content facets is challenging.
- The availability of training data for such content facets often is a challenge - especially if the content facet assessment is addressed via supervised machine learning.
- The common use of Internet slang in social media naturally poses a challenge for Natural Language Processing systems.

Objectivity

Within this dissertation research, the content facet *objectivity* in textual data is defined as follows:

Definition 4.4. Objectivity: Let O be the objectivity of a textual resource R. The textual resource $R = O$ if R does not describe opinions nor private states.

If not otherwise stated, this dissertation refers to this definition of objectivity.

> **Example:** In a blog, one blog post R_{B1} describes the most important facts in respect to iPAD 2, therefore $R_{B1} = O_{objective}$. Another blog posts R_{B2} contains the experiences of an actual user of the iPAD 2, , therefore $R_{B1} = O_{subjective}$. A third blog post R_{B3} contains the blogger's opinion on the Apple company in general. Even though the blog apparently contains some factual information, the whole blog would be classified as being subjective due to the fact that the whole blog expresses more opinions than facts, therefore $R_{B1} = O_{subjective}$.

Challenges The first challenge in the assessment of the content facet objectivity is to be as topic independent as possible and to extract features denoting objectivity characteristics. The second challenge is the availability of training data if objectivity assessment is addressed via supervised classification.

Techniques to assess Objectivity Objectivity/Subjectivity classification has been tackled in the past by a research group around Wiebe and Wilson who also created a gold standard test dataset for subjectivity classification [Wiebe and Bruce, 1999]. This dataset is referred to as the *MPQA corpus* and it is available online[7]. Wilson and Wiebe et al. also released a subjectivity lexicon consisting of a list of subjectivity clues [Wilson et al., 2005] that resulted in a collection of gold standard manual subjectivity sense annotations [Wiebe and Mihalcea, 2006].

In the field of subjectivity classification, the goal is to discriminate between subjective and objective statements using machine learning, classification respectively [Raaijmakers and Kraaij, 2008]. In this work, they exploit character n-grams

[7]MPQA Releases - Corpus and Opinion Recognition System `http://www.cs.pitt.edu/mpqa/`, last accessed Feb 2011

for subjectivity classification with multinomial kernel machines. More specifically, they used bi-, tri-, and quadrigrams whereas in one setting, they considered whitespaces between words as delimiter for an n-gram and in another setting, not. Their experiments revealed that when not considering whitespaces as word n-gram delimiter, an accuracy of 82.5% can be achieved on the MPQA corpus.

In [Remus, 2011], sentence level subjectivity classification on the movie-reviews corpus[8] is improved exploiting several readability measures in combination of features that correspond to strong subjective clues.

Jiang et al. used subjectivity classification in blogs to assess the political leaning of political blogs [Jiang and Argamon, 2008a, Jiang and Argamon, 2008b]. Based one a per sentence classification, they classify blogs into liberal or conservative. Their approach identified subjective sentence that exhibit at least two subjective clues using a subjectivity lexicon. From the subjective sentences, they extracted opinions to create classifiers capturing political leaning features.

In [Huang et al., 2006b], Huang et. al propose a method to categorize blogs by subjectivity using a general linguistic feature, Part-Of-Speech. In their work, especially the problem of "unseen words" - words that never appear in the training data - is tackled in detail. Pang et al. used subjectivity categorization as a preprocessing step for sentiment analysis [Pang and Lee, 2004]. To determine the sentiment polarity, they propose a text categorization technique to just the subjective portions of the document. The extraction of these portions is implemented using efficient techniques for finding minimum cuts in graphs.

Emotionality

In this thesis, the content facet *emotionality* is defined as follows:

Definition 4.5. Emotionality: Let E be the emotionality of a textual resource R. The textual resource $R = E$ if R contains expressions of a person's attitude towards an arbitrary entity or a private state. The emotionality E can be expressed as e.g. joy, anger, or fear.

If not otherwise stated, this dissertation refers to this definition of emotionality.

[8]Movie Review Data, `http://www.cs.cornell.edu/people/pabo/movie-review-data/`, last accessed May 2011

> **Example:** A newspaper article R_N contains an interview with an emotional eyewitness account $E_{emotional}$ of the Tsunami in Japan in 2011. A blog R_B contains information how to react when a tsunami warning has been announced $E_{neutral}$. The newspaper article is therefore emotional $R_N = E_{emotional}$ whereas the blog is neutral $R_B = E_{neutral}$.

Challenges The challenge in the assessment of the content facet emotionality is again to derive a topic independent feature representation. Since emotionality is typically assessed via supervised classification, one challenge is the availability of a sufficiently large amount of training data. Also challenge is the assessment of emotionality in social media - mainly due to the special use of language and grammar in e.g. blogs.

Techniques to assess Emotionality Emotion classification is related to the field of sentiment classification. Sentiment classification is exploited to identify people's feelings towards products, persons, companies, brands, etc. It is especially beneficial in the media domain to investigate the success of companies, products or services. People who use news retrieval frameworks like the earlier proposed APA Labs framework (see Chapter 3) or social media retrieval frameworks like for instance tweetfeel[9] benefit from an automatic classification of search results according to their sentiment into *positive*, *negative*, and *neutral*. Such a classification enables to get an overview of what people think about current events. Figure 4.1 shows an example of the tweetfeel application that automatically classifies Twitter microblogging posts into positive or negative.

Chesley et al. aim to automatically classify blog posts with respect to their sentiment [Chesley et al., 2005]. For this, they first assess whether a blog post is subjective or objective. The resulting subjective blog posts are then further categorized according to their polarity into positive versus negative. For their classification approach, they use both textual as well as linguistic features whereas they take into consideration especially the following Part-of-Speech: first-person pronouns and second-person pronouns, and the number of adjectives and adverbs. The rationale behind that is that typically adjectives and adverbs capture *how something is received*. Therefore, both indicate subjective considerations. The advantage of this

[9]tweetfeel, `http://www.tweetfeel.com/`, last accessed Feb 2011

Figure 4.1: Example: Twitter Sentiment Classification Application TweetFeel - Results for Query Wikileaks

approach is that the used features are both topic as well as genre-independent. Additionally, the exploit external knowledge, namely the Wiktionary dictionary provided by Wikipedia, to assess adjectives that denote sentiment. Note that as classifier, they used a Support Vector Machine based on LibSVM (see Section 2.4.3).

They also investigated whether blogs are different from other genres expressing sentiment and if so, how. They state that blogs exhibit a diverse rhetorical structure when compared to other media.

Many emotion classification approaches are lexicon based. In such approaches, e.g. the ratio of negative to positive words is calculated. For example, Lin et al. [Hsin-Yih Lin et al., 2008] classify Online news articles into reader-emotion categories exploiting an emotion lexicon. In [Yang et al., 2007], emotion classification of blogs is investigated using Support Vector Machines (SVMs) as well as Conditional Random Fields (CRFs). They train both emotion classifiers on sentence level and generalize the results on document level. Besides, they consider the context of the sentences. Their experiments reveal that the CFR classifier outperforms the SVM.

Emotion classification has extensively been tackled in the SemEval challenge 2007[10] in form of an Affective Text Task. In this task, the emotional meaning of

[10]SemEval Challenge 2007, `http://www.cse.unt.edu/~rada/affectivetext/`, last accessed

news headlines had to be classified into a set of pre-defined emotions as for example joy, fear, or surprise. Additionally, the polarity of such emotional meanings had to be assessed in form of binary classes, positive versus negative. The rationale behind the Affective Text challenge was that an affective analysis of news texts is especially beneficial for mining opinions and for market analysis.

The participating systems employed rule-based approaches exploiting external resources as for example SentiWordNet, WordNetAffect, or WordNet; unsupervised as well as supervised methodologies trained on unigrams, or terms derived from emotion lexicons. The evaluations for the Affective Text task revealed that emotion classification is still hard to solve. For instance, the best system, the rule based classifier exploiting external resources, achieved a precision of 48.97% for the emotion *sadness*. For an introduction into the challenge as well as an overview of the approaches and results of the participating systems, refer to [Strapparava and Mihalcea, 2007].

Credibility

Within this dissertation research, the content facet *credibility* in textual data is defined as follows:

Definition 4.6. Credibility: Let C be the credibility of a textual resource R. The textual resource $R = C$ if R is similar to a trustworthy and accurate source in respect to content and publishing behavior.

If not otherwise stated, this dissertation refers to this definition of credibility. It is worth mentioning that this **novel definition of credibility in textual data** has been established within this dissertation research.

> **Example:** Two blogs R_{B1} and R_{B2} describe a political event. The first blog R_{B1} reports that a politician is suspected to have been accepting bribe money but it is clearly outlined that nothing is proven yet. The second blog R_{B2} insults the politician. Clearly, the perceived credibility of the first blog is higher $R_{B1} = C_{high}$ than of the second blog, $R_{B2} = C_{low}$.

Challenges The notation of credibility certainly depends on the application context. The first challenge is consequently the establishment of a credibility definition that is true within a specific application domain. Typically, the credibility of a resource depends on more than one dimension, each encoding an aspect of credibility. For example, the accuracy of information, its verifiability [Klemm et al., 2001b], or its neutrality are indicators for credibility. Therefore, the second challenge is to identify indicators for credibility that are valid and useful within the application domain.

Techniques to assess Credibility Since more and more web-based information serves as the basis for beliefs, decisions, and choices, it is crucial to evaluate the content in respect to credibility and reliability [Harris, 1997]. According to the Stanford Guidelines for Web Credibility[11] one should make it easy to verify the accuracy of the information on Web sites. Especially in the media domain, the availability of credibility scores is beneficial since news definitely serve as decision support for people.

The concept credibility is highly user specific and application dependent. In general, the decision on whether content is credible, depends on multiple dimensions. In [Iding et al., 2008], credibility is defined as *"the ability to inspire belief or trust"* and in [Klemm et al., 2001a], as information accuracy and veracity.

In [Murakami et al., 2010], a methodology is described that supports users in judging the credibility of Japanese Web content. In this approach, a so-called *statement map* [Murakami et al., 2009] is created that shows different viewpoints on a specific information in addition with supporting evidence. More specifically, contradicting, agreeing, conflicting, statements as well as statements that provide confinement as well as evidence are presented to the user. The extraction of these statements is achieved using semantic techniques and Natural Language Processing (NLP).

User studies showed that users judge a Web site credible primarily based on structural and author specific elements [Fogg et al., 2001]. Note that in contrast to standard Web sites, people can be less controlled and are even harder to trace in the blogosphere [Agarwal and Liu, 2008] and other Web 2.0 applications. A blogger, for example, is only identified by the user name, often nicknames, with no meaning.

[11] http://credibility.stanford.edu/guidelines

Because all blogs share a similar structure - the most important are title, date, and content - structural information is less significant for credibility analysis in blogs.

Generally, the credibility of blogs can be derived from e.g. their content, the credibility of the author of the blog, or external references as for example trackbacks or links [Kang, 2010]. Trackbacks and links can be used to create trust and reputation networks, see [Gruhl et al., 2004, Kale et al., 2007, Pujol et al., 2002]. The use of such networks for credibility analysis has been motivated by Web site authority ranking. Kleinberg [Kleinberg, 1999] and Page et al. [Page et al., 1998] developed a ranking mechanism for Web search engines based on the idea of citation networks [Garfield et al., 1984]. Citations in scientific papers are clearly a measure of credibility because published papers are reviewed and therefore no artificial papers exist to boost the citation count. However, citation networks on the Web can be spammed and consequently loose their credibility if not validated.

A popular example for this technique of spamming a citation network on the Web are linkfarms. Linkfarms are an accumulation of Web sites or domains and are used to link as many hyperlinks as possible to another Web site. Linkfarms are used to manipulate search engines ranking techniques as for example the Google PageRank algorithm [Langville and Meyer, 2006]. The PageRank algorithm ranks the relevance of a Web site by the quality and quantity of the sites that link to it. In PageRank each Web page has a measure of authority, the authority of a page is proportional to the sum of the authority scores of pages linking to it. Early versions of the PageRank algorithm have been sensible to link spamming what has been exploited. Nowadays, Google uses about 200 signals to rank search results[12].

The second outlined system is the so called WISDOM system that has been introduced by Akamine et al. in [Akamine et al., 2009]. WISDOM aims to support human judgement on information credibility. The WISDOM system enables the user to search a data set and the search result is analyzed according to several aspects which are related to information credibility. WISDOM uses semantic Natural Language Processing (NLP) techniques to extract the major expressions of a search result. Also, it aims to identify contradictory expressions or opinions in order to provide users with the possibility to judge the credibility of an expression. Additionally, WISDOM provides information about the sender, or author of the content, respectively. Generally speaking, the WISDOM system analyzes the non-topical

[12]http://googleblog.blogspot.com/2008/03/why-data-matters.html

information to a topical search query in detail. The rationale behind that is that the credibility and quality of a search result set more or less depends on **how a topic is dealt with**. Therefore, topic independent aspects of content have to be taken into consideration when judging content quality and credibility. A screenshot of the WISDOM system taken from [Akamine et al., 2009] is shown in Figure 4.2. As a planned feature, the creators of WISDOM aim to incorporate the information

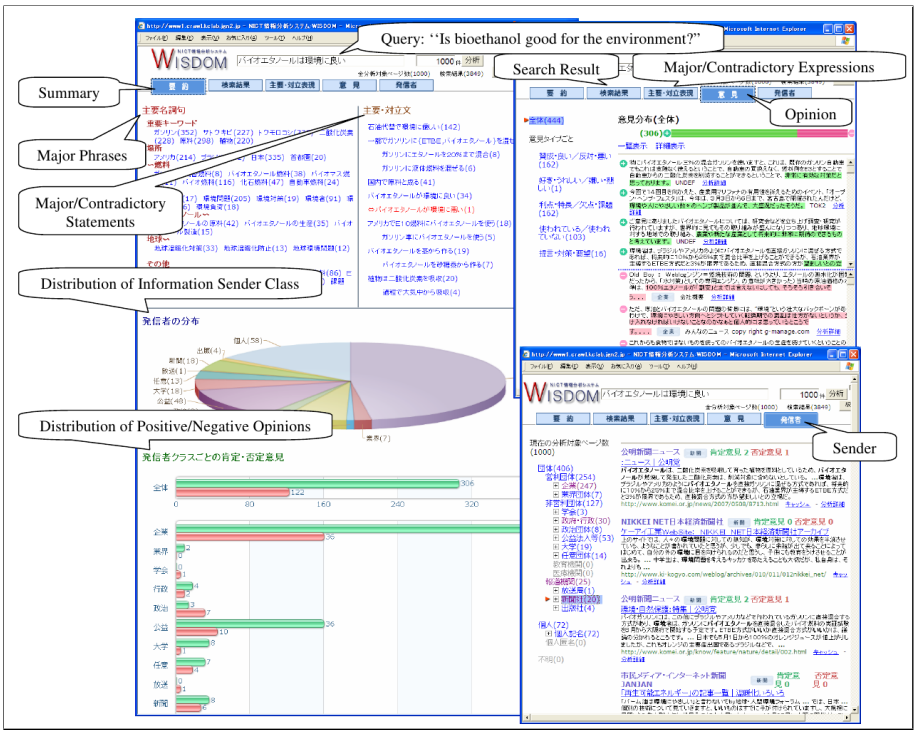


Figure 4.2: The WISDOM Credibility Analysis System

appearance as additional visual indicator of the quality of a Web site. For instance, a large amount of questionable (adult etc.) advertising rather indicates a low perceived credibility of the Web page. The WISDOM system is available online[13]. Note that the system currently only supports Japanese Web sites.

Quality

In this dissertation research, the content facet *quality* in textual data is defined as follows [Eppler, 2003]:

[13]WISDOM `http://wisdom-nict.jp/`, last accessed Feb 2011

Definition 4.7. Quality: Let Q be the quality of a textual resource R. The textual resource $R = Q$ if R is written objectively, neutrally and fact oriented.

If not otherwise stated, this dissertation refers to this definition of quality.

> **Example:** In a blog R_B, a person describes her daily life using Web specific language as for example Web specific acronyms like "rofl" and emoticons like ':-)". This blog R_B would be categorized as a low quality blog, $R_B = Q_{low}$. A news article R_N tackles the advantages and the disadvantages of nuclear power. No opinions or biased statements are expressed. Consequently, this news article would be classified as high quality article, $R_N = Q_{high}$.

Challenges The first challenge is to identify a suitable definition of quality within the application domain. In general, the quality of content strongly depend on the user context. If a person is interested in e.g. Hollywood stars, the tabloid newspaper "The Sun"[14] is a good quality source. Besides, information quality depends on the information context, the domain, respectively.

The second challenge naturally is the availability of training data and the identification of suitable features, if the quality assessment is accomplished using supervised Machine Learning.

The third challenge is to define dimensions encoding information quality in the application domain. It is generally understood that information quality is made up of multiple dimensions [Yoo, 2011]. Clearly, these dimensions can be manifold and definitely, they are rather user specific. For instance, in communities as for example in Yahoo Answers [15], a community based question answer (QA) site by Yahoo!, the quality of an answer is assessed based on user ratings and author network information. In other application scenarios like for instance Facebook microblogging, user rating information may rather indicate popularity than information quality.

Certain measures for information quality are yet commonly agreed upon: for instance, objectivity and accuracy are common indicators for a high level of information quality [Wang and Strong, 1996]. In general, information quality can be

[14] `www.thesun.co.uk`
[15] `http://www.answers.yahoo.com`

divided into four categories: (i) intrinsic information quality, (ii) contextual information quality, (iii) representational information quality, and (iv) accessibility information quality [Katerattanakul and Siau, 1999, Wang and Strong, 1996].

Techniques to assess Quality The intrinsic information quality covers measures like accuracy, objectivity, believability, or reputation [Agichtein et al., 2008]. To assess the intrinsic information quality, external sources can be incorporated to check the accuracy and the correctness of published facets in a piece of information [Magdy and Wanas, 2010]. Besides, the social network of the author can be analyzed to derive the author's reputation or trustworthiness [GiHong and SangKi, 2008, Guha and Tomkins, 2004, Ziegler and Lausen, 2005]. Another approach is to perform a deep content analysis; for instance to assess whether the content is e.g. written objectively [Lex et al., 2010c] and of good style [Blumenstock, 2008].

The contextual information quality is related to relevancy, whether value is added with a piece of information, timeliness, completeness and the amount of information which exist in the context of an information[16]. It can be assessed by comparing a piece of information with a reliable source over time [Juffinger et al., 2009a], or by incorporating external sources as well.

The representational information quality is characterized by the interpretability, ease of understanding, concise representation, and consistent representation of information. It can be assessed via computing e.g. a variety of readability measures [Hoorn, 2010]. Lastly, the accessibility information quality is correlated to access security and accessibility.

Naturally, the discussed information quality dimensions may serve as separate content facets that can be used to support users in e.g. retrieval tasks. For instance, the blog search engine BLOGRANGER[17] [Fujimura et al., 2006] provides a set of search interfaces each addressing a different facet: (i) topic, (ii) blogger (authority of the blog), (iii) links (reputation of the author), and (iv) sentiment (bias of the content). These facets can all be used in a combinatorial search to address **both topical aspects as well as quality aspects** of blogs. For instance, the author

[16]In this context, it is important to mention that in general, information quality should not be confused with information quantity. For example, there exist more pages on the Web reporting that the moon landing didn't happen than pages describing that it actually did happen.

[17]BLOGRANGER, `http://ranger.labs.goo.ne.jp/`, last accessed Feb 2011, available only in Japanese

information can be used to verify her expertise; analyzing the outgoing links might be beneficial since a quality blog should rather link to quality resources. The sentiment information can be used to judge a blog's level of neutrality - which is also a strong indicator of quality [Fujimura et al., 2006].

4.3 Cross-domain Properties of Content Facets

This section provides a discussion of the cross-domain properties of the earlier introduced content facets. In traditional media, typically some content facets are already available. For example, newspapers organize news articles into commonly agreed upon newspaper categories as for example politics, or sports. These categories have been assign by newspaper editors or journalists. Therefore, they can be regarded as high quality labels that can be directly exploited for classification.

However, within this thesis, it has so far been established that traditional media and social media differ in some aspects. Therefore, in this section, it is discussed whether facets that valid and meaningful in traditional media can also be used for social and online media.

Related work in the field of news retrieval in traditional and social media has revealed a strong evidence that the blogosphere correlates with the real world based on quantitative and qualitative analysis. Drezner and Farrel [Drezner and Farrell, 2004] were able to show that there is an interdependency between blogs and the real world.

Topic dependent facets, like topics or even genre can typically be transformed from one domain to the other. Social media provide a discussion to current events and therefore, they deal with similar topics as traditional media; however in a different manner since blogs often express opinions and different perspectives to current events.

Besides, some topic independent facets should be transferable over different media domains. For instance, a date facet is feasible for both traditional media and online news as well as social media. However, one has to take into consideration possible time shifts. Some news are first published in traditional media and then further propagated and discussed over social media, while others are first published over highly dynamic social media before they are published in traditional newspapers. Therefore, if temporal facets should be transferred, a time based alignment has to be performed. See Section 5.5.4 for an implementation of such a temporal

alignment. Note that these temporal drifts pose interesting research question regarding the influence of one type of media onto another. In other words, *who posted first*, or *who influenced whom*.

Note that Section 5.4.2 contains a detailed experimental analysis of the cross-domain properties of a selection of content facets that have been assessed within this thesis.

4.4 Conclusions

This chapter provides an overview of the related literature in the area of the automatic assessment of content facets. A general categorization of content facets into two major categories is presented. As corroborated by related work, content facets can be divided into (i) *topic-oriented* and (ii) *topic-independent* content facets. **This finding provides an answer to Research Question 2.**

This chapter also outlines a variety of content facets whereas the selection has been based on describing content facets that have actually been implemented within this dissertation research. For each of the assessed content facet, a formal definition has been established. To the best of the author's knowledge, such definitions of content facets have not been published so far. Therefore, this thesis adds to the state of the art by providing such formal definitions. Besides, for each content facet, a literature study has been conducted with focus on how to assess the content facet. A key finding has been that for many content facets, there is a strong need to be as topic independent as possible. The literature study also revealed that to assess content facets, not only words are used as features but also classes of words denoting e.g. subjectivity, special punctuations (emoticons), document layout (Web site structure), sentence type (contradicting statement) etc. This finding served as basis for tackling Research Question 3, *Which types of features are most suitable for detecting topic oriented facets and topic independent facets?* Therefore, these insights have been integrated in the selection of the feature sets used for this thesis. This is described in more detail in the next chapter, in Section 5.3.

Table 4.1 gives an overview of the next chapters; more specifically which of the experiments described in Chapter 5 and Chapter 6 contribute to which content facet described in this chapter.

Table 4.1: Overview of Proposed Content Facets and According Experiments

Experiment	**Content Facet**	Section
Keyterm Extraction in News Articles	Topic	3.3
Named Entity Extraction in News Articles	Topic	3.3
Cross Domain Genre Classification from News to Blogs	Genre	5.4.2
Single Domain Genre Classification in Blogs	Genre	5.4.3
Emotion Classification in Blogs	Emotionality	5.5.3
Quality Classification in Online News	Quality	5.5.1
Objectivity Classification in Blogs	Objectivity	5.5.2
Blog Credibility Ranking in News	Credibility	5.5.4
Quality Assessment in Blogs	Quality	5.5.5
Content Facet Assessment in Web Content	Genre and Quality	6.2

Chapter 5

Automatic Assessment of Content Facets

"Where is the life we have lost in living? Where is the wisdom we have lost in knowledge? Where is the knowledge we have lost in information?" (T. S. Eliot)

5.1 Introduction

The aim of this chapter is to outline the experiments that have been conducted to automatically assess content facets in traditional media, online media, and social media.

In this chapter, a feature study is described to derive the best performing features for both topic oriented and topic independent content facets. This provides insights related to Research Question 3, *"Which types of features are most suitable for detecting topic oriented facets and topic independent facets?"*, because the feature study reveals that for topic independent content facets, style based features serve best, while for topic oriented content facets, Bag-of-Words features are best suited. The best features are then used for several content facet classification tasks in the media domain.

The contribution of this part of the thesis is (i) a feature study of style based features and bag of words features, (ii) the assessment of both topic oriented and topic independent content facets, and (iii) the application of selected features and

content facets in the context of an international challenge, namely the TREC Blog Track 2009.

5.2 Methodology

This section briefly outlines the methodology of this dissertation research. Based on this methodology, content facets have been assessed in media content. More specifically, the conducted experiments follow and implement this methodology. The proposed methodology involves the following six steps:

1. Creating a significantly large amount of training data by annotating document corpora
2. Determining and extracting characteristic features for each of the content facets
3. Analyzing the chosen features in terms of importance for the classification task
4. Using the best features for the classification tasks
5. Determining suitable classification algorithms
6. Performing evaluations of the classifier performance using standard evaluation measures

As described by [Wilson, 2008], who proposed a research methodology that is similar in some aspects, such a methodology can be regarded as *"cycle that is often found [...] in research in natural language processing"*. Naturally, at the end of this methodology, the insights gained in the evaluation step can be used to re-run the experiments with e.g. different features or algorithms.

In this dissertation research, Mutual Information and Linear Correlation are used to extract the best performing features for the classification tasks. For evaluation purpose, measures that are standard in Information Retrieval as well as Machine Learning as for example Accuracy, Precision, and Recall are used (see Section 2.4.4).

5.3 Used Content Features

For this dissertation, two categories of content features are proposed and applied to assess content facets: (i) lexical features, and (ii) stylometric features.

5.3.1 Lexical Features

The proposed lexical features are so-called bag of words features. They are derived by first using a Part-of-Speech tagger to tag the documents with the most common Part of Speech (POS). For this dissertation research, as Part-Of-Speech tagger, OpenNLP[1] is used. Then, the particular POS are statistically analyzed in respect to their corpus distribution. Each of the POS statistics is tackled in a separate feature space. The following word statistics have been used:

- *Unigrams:* The number of distinct words that occur in a document
- *Bigrams:* The number of groups of two words that sequentially occur within a sentence
- *Trigrams:* The number of groups of three words that sequentially occur within a sentence
- *Stems:* The number of words that have been stemmed using the Porter Stemming algorithm implemented within the Snowball Stemming Framework[2]
- *Nouns:* The number of all nouns in a document
- *Verbs:* The number of all verbs in a document
- *Adjectives:* The number of all adjectives and adverbs in a document
- *Leading Graphem:* The number of character n-grams of size three leading any token in the document
- *Trailing Graphem:* The number of character n-grams of size three trailing any token in the document

[1] `http://opennlp.sourceforge.net/`
[2] `http://snowball.tartarus.org/`

- *Personal Pronouns:* The number of all distinct personal pronouns that occur within a document

The advantage of lexical features is that they are simple and easy to extract, as has also been stated in [Karlgren and Cutting, 1994].

5.3.2 Stylometric Features

For this dissertation research, a number of stylometric features has been selected with focus on topic independence. Since topics change rapidly in the media domain, it is crucial to create classifiers that are independent of topics. The following topic independent stylometric features have been used whereas they have firstly been introduced in [Lex et al., 2010a]:

- *Punctuation:* The punctuation distribution that is defined as the count of one of 12 punctuations per document
- *Emoticons:* The average number of sequential double, triple, and n punctuations
- *Words in sentences:* The distribution of sentences with different word lengths as for example 0-3,4-6,7-9,10-12,13-15,...
- *Average Number of Words / Sentences:* The average number of words per sentence
- *Characters in Sentences:* The distribution of sentences with a number of 0-20,21-40,41-60,61-80,81-100,... characters
- *Average Number of Characters / Sentences:* The average number of characters per sentence
- *Noun+Verb Sentences:* The average number of minimal (in-)correct sentences whereas the sentence correctness is defined so that a correct sentence must at least have a noun and a verb
- *Average Number of Unique POS Tags:* The average number of unique Part of Speech tags per sentence

- *Lower Case/Upper Case:* The ratio of lower case characters to upper case characters in a document
- *Word Length:* The word length distribution, number of words of length 1,2,3...8 and the average word length in a document
- *Adjective Rate:* The number of adjectives divided by the number of all tokens of a document
- *Adverb Rate:* The number of adverbs divided by the number of all tokens of a document

Since style does not depend on topics, stylometric features are inherently topic independent. Therefore, they provide a high degree of generalizability in an inhomogeneous topic landscape like the media domain while being simple.

5.4 Experiments: Topic Oriented Content Facet Assessment

This section describes the experiments that have been conducted to assess topic oriented content facets in the media domain.

The methodology proposed in Section 5.2 involves supervised classification; therefore, training data has to be created beforehand. In traditional media, news articles are typically already labeled: Newspaper editors assign the news articles to commonly agreed upon newspaper genres like e.g. politics or to popular topics like e.g. Olympic Games. In contrast to that, in social media, such high quality labels[3] are lacking. Even though blogs may be labeled, their tagging vocabulary is rather heterogenous and not commonly agreed upon.

This section provides insights related to Research Question 1, *"How effectively can classification schemes from traditional media be mapped onto blogs? "*, because this thesis proposes to exploit the high quality labeled training data from the news domain to classify blogs that also deal with news related events. In order to investigate the feasibility of this idea, firstly, it has been investigated whether there is a

[3]The labels are of high quality because domain experts, that is journalists or newspaper editors assign them manually to the news articles before they are published.

correlation between the content of traditional news articles and of blogs. In other words, the *collective opinion* in social media is compared against the *news opinion*.

5.4.1 Content Correlation Analysis

The content correlation analysis experiment aims at analyzing the content wise correlation between news articles and news related blogs [Juffinger and Lex, 2009]. For the content correlation analysis experiment, the Pearson's product moment correlation coefficient[4] has been used.

Dataset

For this experiment, around 40 blogs have been manually selected whereas special attention has been paid to guarantee that the blogs actually deal with news related events. The blogs have been selected based on their popularity, their actuality, and their significance. This has guaranteed that the blogs are actively maintained and deal with current events. Since the blogosphere is multilingual, the blog selection included English, German, and French blogs. Note that the blogs have been equally distributed over these three languages.

The news corpus consisted of German news articles from the news repository maintained by the Austrian Press Agency (APA)[5]. The news articles have been crawled in a time period of two months in 2009.

Approach

The content correlation between news and blogs has been evaluated in respect to the following two aspects:

1. Language
2. Query Term Type

For the query term type aspect, 15 person names, 15 locations, as well as 15 arbitrary query terms have been used, summing up to 45 queries. The language

[4]Wolfram Mathworld, Correlation Coefficient, `http://mathworld.wolfram.com/CorrelationCoefficient.html`, last accessed May 16, 2011

[5]www.apa.at

aspect involves a computation of the correlation between German news articles and French, English, and German blogs. Note that in this case, a cross-language retrieval step has been applied to translate the query terms into the target language. Note that the cross-language retrieval procedure has been the work of Juffinger et al. [Juffinger et al., 2009b].

The procedure of the content correlation analysis is described in terms of pseudocode in Algorithm 1:

Data: Two sets R_1, R_2 of search results $r_1, .., r_N$ retrieved from blogs and news corpora for the query q of a type T. The sets R_1 and R_2 are represented as time series and ordered in ascending time. Multiple results per time entry r_i are possible.
Result: The correlation between R_1 and R_2.

foreach r_i *in* R_1 **do**
  $newsResults[i]$ = `#QueryResults(`r_i`)` ;
end
foreach r_i *in* R_2 **do**
  $blogResults[i]$ = `#QueryResults(`r_i`)` ;
end
correlation = `pearsonsCorrellation` (newsResults,blogResults);

Algorithm 1: Content Correlation Analysis

This procedure is carried out for each of the three languages, resulting in a correlation between news and blogs for German, French, and English.

Approach

The results for the content correlation analysis are given in form of box plots. There is one boxplot for each query type and each box plot shows the results for all three languages. Generally, the box plots depict minimum and maximum of the correlation between the time series of news and blogs, as well as the standard deviation and the mean.

Figure 5.1 shows the correlations between the German news articles and German (DE), English (EN), and French (FR) blogs. In this case, the query term type has been *Persons*. This experiment reveals that the system performs similar for German and French; yet the correlation between German news and English blogs is lower.

Figure 5.2 shows the correlations between the German news articles and German

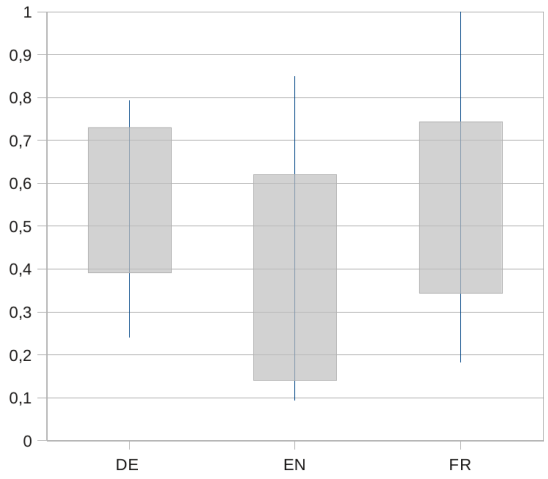


Figure 5.1: Correlation between News and Blogs for Query Type Persons

(DE), English (EN), and French (FR) blogs. In this case, the query term type has been *Locations*. This experiment reveals that the system performs similar for all

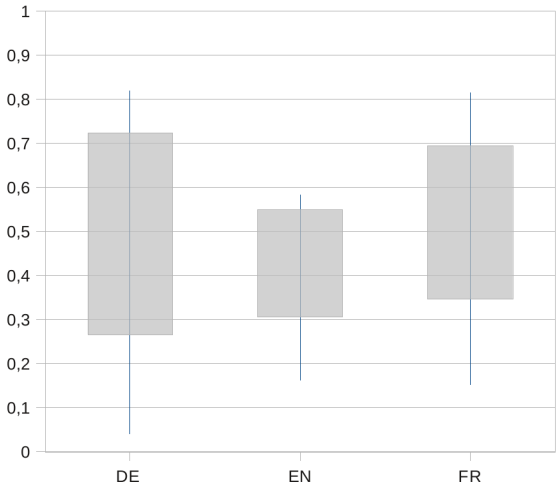


Figure 5.2: Correlation between News and Blogs for Query Type Locations

three languages whereas the standard deviation for German blogs is higher.

Figure 5.2 shows the correlations between the German news articles and German (DE), English (EN), and French (FR) blogs. In this case, the query term type has been arbitrary *Terms*. In this case, the correlations between the languages are quite similar as well; this can be interpreted so that the system is robust for arbitrary query terms.

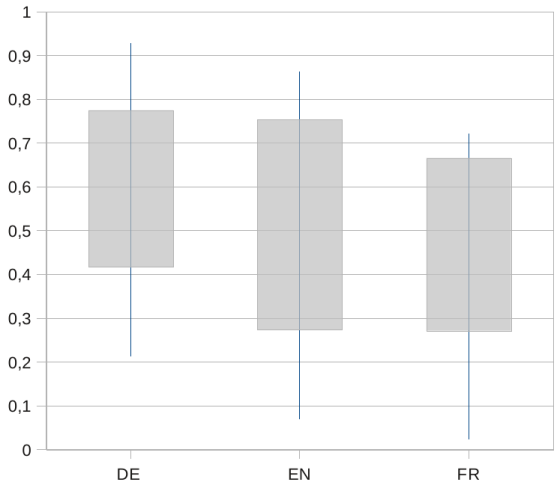


Figure 5.3: Correlation between News and Blogs for Query Type Terms

The overall correlation coefficient for the cross-language content correlation analysis has been 0.55 with a standard deviation of 0.18, a maximum of 0.93 and minimum of 0.02. Therefore, one can draw the conclusion that the system is robust for the different types of query terms and languages.

Lessons Learned

The content correlation analysis experiments has revealed that traditional news articles and selected blogs are correlated over time across different languages in respect to person, locations, as well as arbitrary query terms. Therefore, the proposed idea to infer knowledge from traditional news to blogs is feasible. In the next section, this insight is exploited to use newspaper categories from traditional news as training data to categorize news related blogs.

5.4.2 Cross Domain Genre Classification from News to Blogs

To further investigate whether news topics are propagated to the blogosphere, a cross-domain genre classification experiment has been conducted. For this, blogs have been classified into commonly agreed upon newspaper genres [Lex et al., 2009c, Lex et al., 2009b, Lex et al., 2010e] following the formal description of the content facet genre in Section 4.2.2.

In the media domain, news articles are typically carefully assigned to predefined

newspaper categories by newspaper editors; the question is: can these annotated news articles be exploited as training data to classify news related blogs?

When no training data available for a domain, it is possible to train on an akin domain and use the resulting training data to label documents from another domain. This is referred to as *Cross Domain Classification* [Xue et al., 2008].

For the cross-domain genre classification experiment, two corpora have been created: a German news corpus [Lex et al., 2008] and a blog corpus that contains German news related blogs [Juffinger et al., 2009a]. The newspaper corpus has been manually labeled by newspaper editors in five common newspaper categories: politics, economy, sports, culture and science. The blog corpus has been crawled from the World Wide Web.

For this experiment, several text classification algorithms have been applied and evaluated in terms of the classifier performance. The classification task is described in more detail in the next section.

Classification Task

The classification task consists of a cross-domain multi-class problem with five classes. The classes correspond to the selected newspaper categories, politics, economy, sports, culture and science.

Three common text classification algorithms have been applied for this experiment: (i) the Class-Feature-Centroid (CFC) [Guan et al., 2009] classifier is used (see Section 2.4.3), (ii) a k-Nearest Neighbor (k-NN) algorithm (see Section 2.4.3), and (iii) a Support Vector Machine (SVM) based on LibLinear [Fan et al., 2008] (see Section 2.4.3).

Experimental Settings

To measure the cross-domain performance of the three classifiers, the following four scenarios have been evaluated:

NN NewsNews: The training set of the news corpus has been used to train the classifiers and the performance on the news evaluation set is reported.

BB BlogBlog: The training set of the blog corpus has been used to train the classifiers and the performance on the blog evaluation set is reported.

NB NewsBlog: The training set of the news corpus has been used to train the classifiers and the performance on the blog evaluation set is reported.

BN BlogNews: The training set of the blog corpus has been used to train the classifiers and the performance on the news evaluation set is reported.

NB-B NewsBlog-Blogs: The classifiers are trained on both the news and the blog corpus. The classifier performance has been measured on a blog evaluation set that was not part of the training set.

NB-N NewsBlog-Blogs: The classifiers are trained on both the news and the blog corpus. The classifier performance has been measured on a news evaluation set that was not part of the training set.

From these different settings, statements have been derived about the generalization ability of all classifiers. The assumption is that the settings NN and BB exhibit the best results because they are not cross-domain tasks. Therefore these settings have served as a baseline for the cross-domain tasks. Also, training has been performed on a combination of both corpora injecting different amounts of documents from the particular target domain (settings NB-B and NB-N). The assumption is that classifiers trained on a mixture of both news and blogs documents are more accurate because they are able to better capture the vocabulary of both datasets. From these experiments, comparisons can be made with the results from NB and BB. The results of these experiments are described in Section 5.4.2.

In order to investigate the applicability of the cross-domain classification models further, both the news corpus and the blog corpus have been subject to a term-based statistical analysis. The goal of the statistical analysis has been to investigate to what extend the term distributions of both corpora differ from each other (see Section 5.4.2).

In the following, the used datasets are described in more detail.

Dataset Properties

The first corpus, the news corpus, has been created from the news repositories maintained by the Austrian Press Agency (APA). Since APA mainly publishes German news articles, the created news corpus contains only German articles. The

news corpus contains around 28k documents whereas for each of the five newspaper categories, approximately the same number of documents is available (∼5600 documents). In other words, the news corpus is balanced.

For the experiments, the news corpus has been Part-of-Speech (POS) tagged. The goal of this POS tagging step has been to extract all nouns from the documents. The rationale behind this is that nouns typically capture the topical information of documents. It has been established that genre also depends on topics, in this experiments, only nouns are used as features. Note that the POS tagging step resulted in about 237k nouns overall with 92.5 nouns per document on average.

The second corpus, the blog corpus, has been crawled from the Web in the course of the APALabs Blog Trend Visualization project (see Section 3.3). The blog corpus contains around 11k blog entries from 56 German blogs. The blogs have been selected according to the five used newspaper sections. The corpus distribution of the blog corpus is given in Table 5.1.

Table 5.1: Description of the Blog Corpus Distribution for the Cross-Domain Genre Classification Experiment

Category	**# Blogs**	**# Blog Posts**
Politics	10	$\tilde{2}800$
Economy	10	$\tilde{2}800$
Sports	10	$\tilde{2}400$
Culture	11	$\tilde{1}400$
Science	15	$\tilde{1}100$

Note that the blog entries have been labeled with the newspaper category of the whole blog. A selection of blog entries has been randomly checked in respect to the correctness of the assigned label. Since not all blog entry labels have been investigated, the dataset may contain mis-labeled data. Naturally, this limits the theoretically achievable classifier accuracy to less than 100%.

The blog corpus has also been POS tagged which resulted in about 110k nouns with an average document length of 61.5 nouns and a total nouns count of 675k.

Since the goal of this experiment has been to transfer a classification model trained on nouns in the news domain to the blog domain, the noun wise overlap of both the news and the blog corpus has been investigated. For this, the dictionaries of both corpora have been merged. The sum of the news and the blog dictionary is 347k

nouns. The merged dictionary then exhibits 302k different nouns. **Consequently, both corpora share only 45k terms.**

Statistical Analysis

In order to investigate the difference between both corpora further, the statistical difference between the news and blog corpus has been calculated. As measure for the statistical difference, the *Kullback-Leibler divergence* (KL) has been used.

The Kullback-Leibler divergence [Kullback and Leibler, 1951] measures the difference between two probability distributions B and N. The KL divergence between the two corpora (Blog B, News N) is derived in Equation 5.1.

$$KL(B||N) = \sum_{t} \left[P_B(t) log \left(\frac{P_B(t)}{P_N(t)} \right) \right] \tag{5.1}$$

Note that $P_B(t)$ denotes the probability of the term t in corpus B and $P_N(t)$ the probability in corpus N. The results for the Kullback-Leibler divergence for both corpora are shown in Table 5.2.

Table 5.2: Kullback Leibler Divergence between Blog and News Corpus in Cross-Domain Genre Classification Experiment

	BlogNews	NewsBlog	Mean
KL Global	0.535	0.430	0.483

The Kullback-Leibler divergence of 0.535 reveals significant differences in the term distributions. This is corroborated by the earlier finding that both corpora share only a relatively small amount of nouns (only 45k out of 302k terms). Consequently, both corpora are statistically different.

In the next section, the parameter settings for the applied three classification algorithms are described.

Parameter Settings

To weight the document vectors, the BM25 term weighting scheme [Jones et al., 2000] has been used for k-NN and the SVM with the standard parameters $k = 2$ and $b = 0.75$. Also, variants of the term weighting scheme TF-IDF have been evaluated, yet the k-NN and the SVM algorithm performed best with BM25 with standard

parameterization of $k = 2$ and $b = 0.75$. For the CFC algorithm, as recommended by the authors, a standard TF-IDF weighting has been used. For the Class-Feature-Centroid (CFC) algorithm, a parameter study has been conducted to identify the best value for parameter b in CFC. Different from findings in the publication of CFC, where has been set to $b = e - 1.7$, for this experiment, a value of $b = e - 1.0$ served best. The parameter k of the k-NN algorithm has been varied from 5, 10 to 15 in a parameter study and as a result, k has been set to $k = 15$. For the SVM, a L2-loss SVM has been used that has been parameterized with standard values (the cost parameter has been $C = 1$).

The next section describes and discusses the results of the cross-domain genre classification experiments.

Results and Discussion

First of all, the performance of all three classifiers on the single domain classification tasks has been evaluated in terms of classification accuracy. The goal of this performance evaluation has been to derive the maximum achievable classifier performance. Note that the single domain tasks correspond to the scenarios NN and BB. In the next step, the cross-domain scenarios (NB, BN) have been evaluated. Figure 5.4 depicts the classification results in terms of mean classification accuracy. As can be derived from Figure 5.4, the SVM achieves the best results for scenario

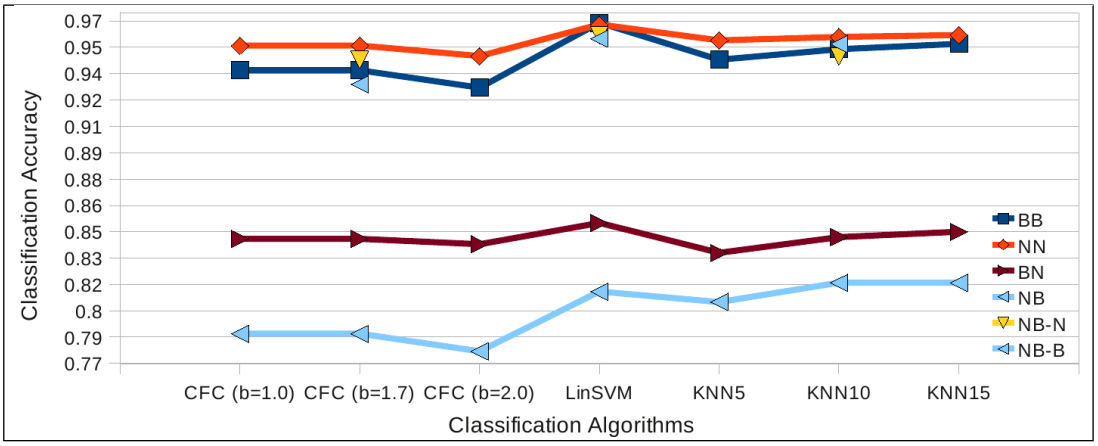


Figure 5.4: Evaluation Results for all Scenarios for the Cross Domain Genre Classification from News to Blogs

BN, the k-NN is second and the CFC algorithm performs slightly worse than the

k-NN algorithm. For scenario NB, the k-NN performs slightly better than the SVM and the CFC.

The results show that in many cases, the SVM performs best and the k-NN slightly better than the CFC, although all accuracy values are within a range of a standard deviation of 0.01.

In respect to computation time, the CFC is by far the best. The training and testing times for all three algorithms are shown in in Table 5.3.

Table 5.3: Training and Test Time of all Classification Algorithms in the Cross-Domain Genre Classification Experiment

algorithm	t_{train} (mean, std dev)	t_{test} (mean, std dev)
k-NN	4.6s / 0.2s	123s / 0.9s
SVM	37s / 0.3s	0.18s / 0.003s
CFC	9.982s / 0.3s	0.197s / 0.1s

As last scenarios, NB-N and NB-B have been evaluated. In these scenarios, documents from the target domain have been step-wise added to the training set. For instance, it has been trained on news and additionally, on 100 blog posts. Then, the classifier performance has been evaluated on the remaining blog posts. The goal of these experiments has been to derive whether it is feasible to annotate at least a subset of documents from the target domain. The rationale behind that is that a classifier should perform better if it has been trained on at least a part of the vocabulary of the target domain. The results for these two scenarios are depicted in Figure 5.5.

In the left figure, the amount of blogs is increased whereas in the right figure the amount of news articles is increased. From this experiment, it can be seen that about 200 blog posts are needed to improve the classifier accuracy. Besides, the CFC algorithm apparently requires a smaller amount of labeled documents from the target domain to increase the overall classifier accuracy. For example, the CFC performance increases with 6%, opposite to SVM (4%) and k-NN (2%) when 200 labeled blog posts are added to the news training data.

To deeper analyze the decisions of the CFC algorithm, the class centroids of the CFC have been subject to another statistical evaluation. For this, only terms with a weight $w > 0$ are considered. As mentioned in the description of the CFC algorithm in Section 2.4.3, these terms are claimed to be the most discriminative terms in the

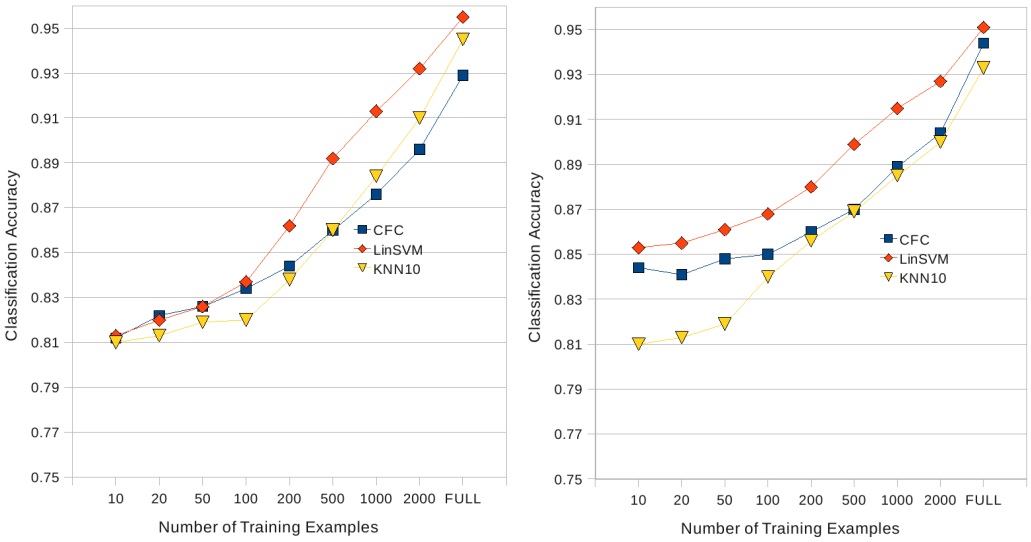


Figure 5.5: Evaluation Results for Scenarios NB-B and NB-B

classes.

Again, the Kullback-Leibler divergence has been computed between the news and blog corpus, but in this case, only centroid terms have been taken into consideration. As shown in Table 5.4, the KL divergence significantly decreases (by a factor 4 on average). These experiments reveal that the CFC actually selects the most

Table 5.4: Kullback-Leibler Divergence of the CFC centroids in the Cross-Domain Genre Classification Experiment

	BN	Std.Dev.	NB	Std.Dev
KL Local Politics	0.139	0.002	0.202	0.003
KL Local Economy	0.218	0.002	0.146	0.009
KL Local Sports	0.132	0.002	0.134	0.002
KL Local Culture	0.096	0.003	0.075	0.002
KL Local Science	0.049	0.001	0.093	0.008

discriminant terms that are characteristic for a class and these remain the same across both corpora.

Conclusions

In this cross domain genre classification experiment, three classifiers have been applied and evaluated on two single-domain and two cross-domain scenarios. The

experiments revealed that the CFC classifier performs comparably with SVM and k-NN while being remarkably faster. Also, in terms of memory complexity the CFC outperforms SVM and k-NN. In the genre classification experiment, the CFC only has to store five centroid vectors - one for each class, whereas the SVM has to store about 1000 support vectors, and the k-NN the full training set (17k for scenario NN and NB).

The experiment showed that newspaper categories, genres, respectively, can be mapped from traditional news to blogs. Therefore, **this experiment provides an answer Research Question 1.**

5.4.3 Single Domain Genre Classification in Blogs

In the previous cross-domain experiment, news classification schemes have been mapped to news related blogs. The news related blogs have thereby been selected manually. However, for large real-world settings, a manual selection of news related blogs is not feasible.

The automatic selection of news related blogs within the blogosphere is challenging. Within this dissertation, this challenge has been addressed with a genre detection related approach [Lex et al., 2010b]. In this approach, blogs have been classified into the genre *news* based on binary classification strategy following the formal description of the content facet genre in Section 4.2.2. More specifically, blogs have been classified into news related blogs versus rest. Note that in the following, this is further referred to as *News versus Rest (NvR)* task.

In the next section, the used dataset and features are described in more detail.

Dataset and Features

The dataset used for the single-domain genre classification in blogs consists of a randomly selected subset of the TREC Blogs08 Dataset[6].

Due to the fact that the TREC Blogs08 dataset does not contain any labeling information, a suitable annotated corpus had to be created by hand.

In the course of this dissertation research, a subset of the TREC Blogs08 has been manually annotated. More precisely, 83 blogs with a total number of 12844 distinct blog entries have been annotated into two classes: *News* and *Non-News*.

[6] `http://ir.dcs.gla.ac.uk/test_collections/blogs08info.html`

Figure 5.5 shows the corpus distribution for this task whereas the distribution of the classes is given both on blog and blog post level. Note that blog level corresponds to summarizing a blog's single posts into one blog document.

	News	Non-News
Blog Level	29%	71%
Blog Post Level	30%	70%

Table 5.5: Blog Corpus Distributions for Single-Domain Genre Classification

From this table, it can be derived that the corpus is unbalanced and biased towards the negative class.

Algorithms

As text classifiers, a Support Vector Machine (SVM) has been used, more specifically, both an SVM based on LibLinear as well as an SVM based on LibSVM, a k-NN algorithm with $k = 10$, and the Class Feature Centroid (CFC) with $b = 1.1$. Additionally, for this experiment, a Naive Bayes classifier, with and without boosting (AdaBoost) as well as a C.45 decision tree with and without boosting (AdaBoost) have been used to especially address the dense stylometric feature space. Note that the Naive Bayes classifier, the AdaBoost, and the C.45 decision tree have all been taken from the Mallet[7] toolkit.

Experiments and Results

For this experiment, the performance of the earlier described text classification algorithms with lexical as well as stylometric features has been evaluated. Firstly, the NvR task has been addressed with lexical features. Three classifiers, namely SVM, k-NN, and CFC have been trained on various lexical features, and the performance of the classifiers is reported both on blog as well as blog post level. Note that the experiments have been conducted using a 10-fold cross-validation. Figure 5.6 gives the results on blog level: The best results are achieved with k-NN on stems (81.3% classification accuracy). Figure 5.7 gives the results achieved with lexical features on blog post level: In this case, the best accuracy is achieved with SVM (LibLinear) also on stems: 91.2%. Consequently, it makes a difference whether the classification

[7] `http://mallet.cs.umass.edu/`

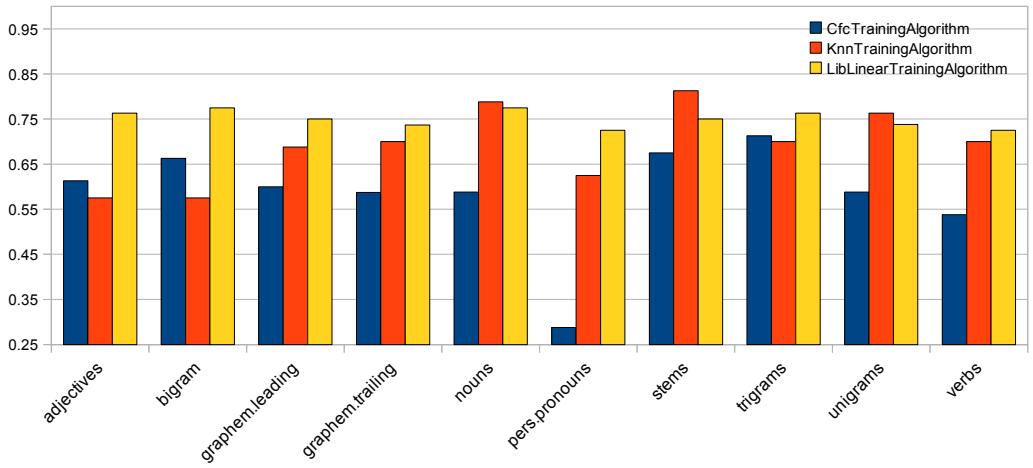


Figure 5.6: NvR Task: Classifier Accuracy of CFC, k-NN, and SVM on Blog Level with Lexical Features

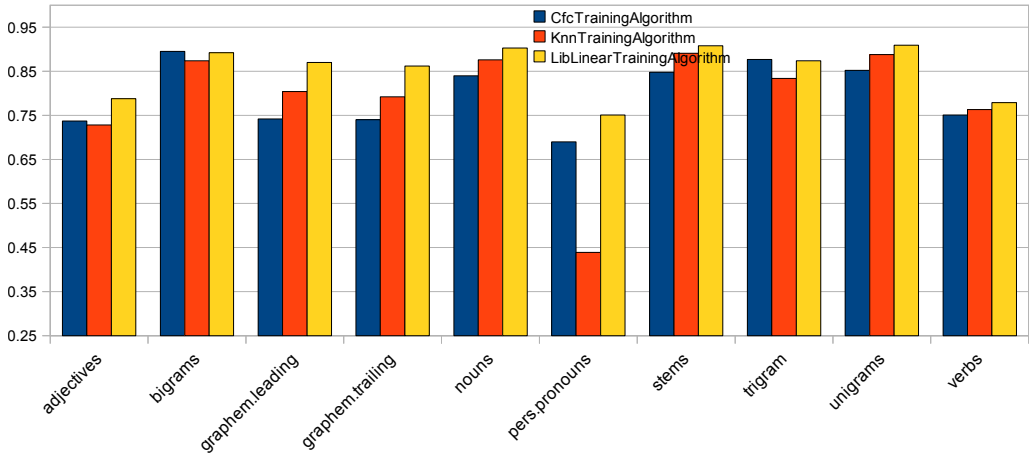


Figure 5.7: NvR Task: Classifier Accuracy of CFC, k-NN, and SVM on Blog Post Level with Lexical Features

is performed on blog or on blog post level. Apparently, the summation of a blog's post into a single blog document performs about 10% worse; this leads to the following conclusion: the assessment of blog genre has to be done on blog post level. Then, the blog post level results can be extrapolated to the blog level.

From an algorithmic point of view, the experiments with lexical features showed that the CFC algorithm performs better as the size of the feature space grows. CFC performs slightly worse on the unigram space and similar to the SVM on the bigram and trigram feature space. Note that the unigram feature space has about 82k dimension, the bigram space about 680k dimensions, and the trigram space 1.42 million. Consequently, the advantages of the CFC algorithm in extremely high dimensional spaces are two-fold: firstly, the algorithm performs better the more dimensions are in the feature space and secondly, the algorithm is extremely fast in terms of training and classification, see Table 5.6. As shown in this table, the

Algorithm	Train(s)	StdDev.	Test(s)	StdDev.
CFC	5.494	1.061	0.037	0.002
KNN	0.034	0.000	63.448	1.078
LibLinear	38.089	1.411	0.036	0.002

Table 5.6: Train and Test Time for CFC on Trigram Features in the Single-Domain Genre Classification Experiment

CFC training is about 10 times faster than the LibLinear training phase and the classification phase outperforms KNN by a factor of 200.

The insight that blog genre assessment has to be performed on blog post level has further been used in the experiments with stylometric features; more specifically, with stylometric features, blog genre has been assessed on blog post level only.

To identify the most relevant stylometric features for blog genre assessment, the Linear Correlation (LC) has been computed: Figure 5.8 shows the LC of the top 20 stylometric features. The LC reveals that the *adjectives/token* feature is the most correlated with about 0.36. The feature *adjectives + adverbs / tokens* also has a very high correlation due to the linear dependency on the first feature. Besides, the sentence length, and the sentence complexity (number of words per sentence) are highly correlated features for the NvR task.

For further computations with stylometric features, only the features with the highest LC have been taken into account. This resulted in features as for exam-

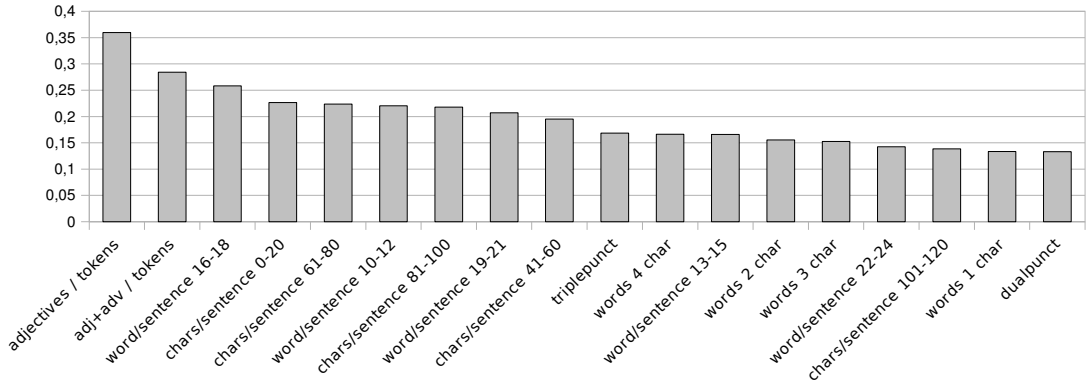


Figure 5.8: Results for Linear Correlation in Single-Domain Genre Classification with Stylometric Features

ple the number of adjectives per token, the number of distinct words per sentence, the average word length, and the emoticon count. These features have then been evaluated using again CFC, LibLinear, and k-NN. Additionally, LibSVM, Naive Bayes with and without Boosting as well as C.45 with and without boosting have been used. Figure 5.9 depicts the results achieved with stylometric features. The conducted experiments with 10-fold cross-validation revealed that the best classification accuracy has been achieved with k-NN ($k = 10$), namely 75%. Due to the fact that the corpus is unbalanced and rather biased towards the negative class, one can draw the conclusion that the classification with stylometric features has not been successful in this case.

A grid search with the LibSVM to determine the best performing cost parameter C resulted in a performance increase of 2%. Nevertheless, the LibSVM performs worse than the other algorithms. A reason for this is that no feature normalization has been performed and the stylometric features are at different scale. This may also be a reason why the k-NN algorithm serves best.

Summary

In this experiment, lexical and stylometric features have been investigated to determine the best performing features and classifiers for the news versus rest (NvR). The aim of the NvR task is to assign blogs to the news genre. To identify the most relevant stylometric features for this single domain genre classification task, the linear correlation has been computed. The conducted experiments revealed that lexical

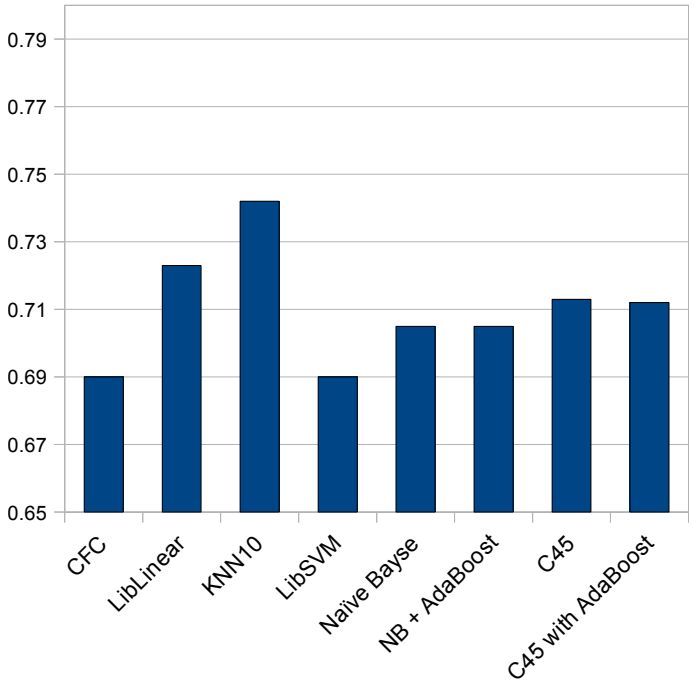


Figure 5.9: NvR Task: Classification Accuracy of CFC, SVM, k-NN, Naive Bayes, C.45 on Blog Post Level with Stylometric Features

features are better suited for the NvR task than stylometric features. The experiments also showed that blog genre has to be assessed on blog post level. Finally, it has been shown that the CFC algorithm performs equally good as SVMs in high dimensional spaces (greater than 1 mill. dimensions), but outperforms LibLinear in terms of time consumption.

5.4.4 Lessons Learned

In this section, a selection of experiments has been presented that were related to topic oriented content facet assessment. The experiments revealed that topics from traditional media are reflected in social media.

In a cross-domain genre classification experiment it has been shown that classification schemes from traditional media can be transferred to social media. **This provides an answer to Research Question 1.**

Additionally, genre classification has been performed on blogs only. In this experiment, the two proposed types of features have been applied: lexical and stylometric features. The experiment revealed that with lexical features, a high classification

accuracy can be reached for genre classification in blogs.

However, since it is explicitly written in the literature, that topics should not be confused with genres, topic independent stylometric features were also applied in the context of genre classification.

The results achieved with this evaluation have been just a little bit higher than the trivial acceptor would have achieved. From this one of the hypothesis of this dissertation research - that genres are not completely independent of topics - is corroborated since for topic independent genre classification, still a feasible methodology is lacking.

5.5 Experiments: Topic Independent Content Facet Assessment

In this section, the experiments are outlined that have been conducted in the context of topic independent content facet assessment.

Several topic independent content facets have been assessed within this dissertation research. Note that the topic independent facets all cover aspects useful for the assessment of information quality. In the following these content facets and their assessment is described in more detail.

5.5.1 Quality Classification in Online News

Within this experiment, the quality of Online news media has been assessed following the formal description of the content facet quality in Section 4.2.3. For this, established Online news has been subject to a binary classification whereas the classes have been *high quality news content* versus *Yellow Press news content* [Lex et al., 2010c]. Note that Yellow Press is a type of journalism that often features exaggerations of news events and sensationalism. Therefore, Yellow Press news typically exhibits rather low perceived quality.

For this experiment, the earlier described stylometric features as well as lexical feature have been applied and evaluated. By means of counter examples, it has been shown that lexical features are not well suited to assess the quality of Online news. The reason for this is with the lexical features, text classifiers implicitly learn topics.

Similarly as in the earlier described single domain genre classification experiment (see Section 5.4.3), the goal has been to determine which features are most applicable for the quality classification task.

Dataset

This section describes the dataset that has been used for this experiment. The dataset consists of Online news articles datasets have been created. The news corpus has been crawled from the Web in 2010 over a time period of approximately three months. It contains articles from British newspapers, two high quality newspapers, namely The Telegraph[8] and The Guardian[9]) as well as articles from Yellow Press newspapers, namely The Daily Mail[10] and The Sun[11]). As mentioned earlier, the Yellow Press newspapers exhibit a rather low perceived objectivity. Summing up, the corpus consists of about 5500 news articles: 3000 high quality articles, and 2500 articles from Yellow Press.

The advantage of these newspapers is that all newspaper articles are tagged by newspaper editors with categorization information. This categorization information denotes the according newspaper categories as for example sports, politics, or science. The tagging information can directly be derived from the news articles' URLs.

For the news corpus, the following tags have been selected from High Quality and Yellow Press: Columnist, Royals, The Royal Family, Diana, Music, Film, Celebrity News, and Bollywood. Note that special attention has been paid to only use categories that have been present in both types of newspapers.

Algorithms

As classifiers, a Support Vector Machine (SVM) based on LibLinear has been used with standard parameterization, a k-Nearest Neighbor algorithm (KNN) with different values for k, and the Class-Feature-Centroid classifier (CFC) with b=1.1.

[8] www.telegraph.co.uk/
[9] www.guardian.co.uk/
[10] www.dailymail.co.uk/
[11] www.thesun.co.uk/

Approach

The approach is based on two steps: Firstly, a uni-domain classification setting has been evaluated. In this context, single-domain means that training and test set consist of documents from all given newspaper categories. In other words, the classifiers are trained and evaluated on the same topics. Secondly, a cross-domain classification setting has been designed in order to investigate the topic (in-)dependency of lexical and stylometric features. For this, a special training set has been created with no topic wise overlap between training and test set.

Experiments and Results

Firstly, a set of lexical features has been applied to the uni-domain classification setting. In this task, the classifiers have been trained on all topics. Then, the lexical features have been applied to the cross-domain classification setting. The results of these two experiments are shown in Figure 5.10. Note that a 10-fold cross-validation has been used. In the uni-domain setting, all classification algorithms achieve very

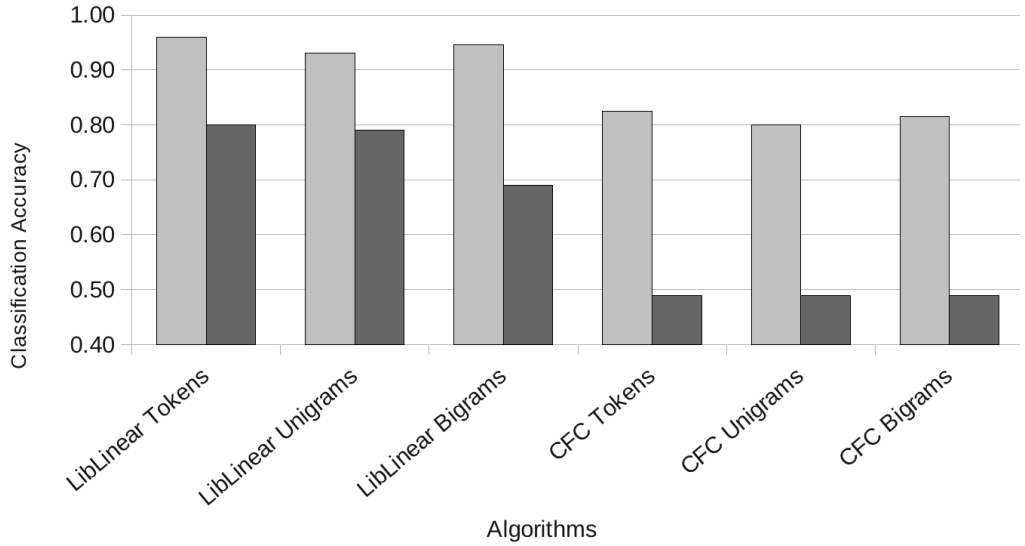


Figure 5.10: Performance on News: Classification Accuracy of SVM and CFC with Lexical Features in Uni-Domain and Cross-Domain Setting

good results. The worse performance for the unigrams and bigram features reveal that the classification works better when the words are not stemmed and stopwords

are not removed. Apparently, stopwords as for example pronouns carry valuable information for quality classification.

In the cross-domain setting, the classifier performance of all algorithms significantly drops. Since there is no topic wise overlap between training and test set in this case, this reveals that lexical features inherently learn topics. Consequently, if the classifiers are trained on completely different topics than they are tested on, the algorithm performance is lower.

The CFC algorithm achieves the worst results. The reason for this is its weighting scheme. The weighting scheme generally assigns discriminative terms a higher weight. Clearly, the CFC selects topic terms as discriminative terms between the classes in the training set.

After having evaluated the lexical features in both the uni- and and the cross-domain setting, the stylometric features have been evaluated. In order to derive the best performing features, first the Mutual Information (MI) has been computed between the stylometric features and the class labels. The features with the highest MI are listed here:

- *Average Number of POS tags / Sentence STD:* Standard deviation of average number of POS tags per sentence
- *Average Word Length:* Denotes the average word length.
- *Subjectivity / Token:* Ratio of subjective words from subjectivity lexicon to all tokens
- *Objectivity / Token:* Ratio of objective words from subjectivity lexicon to all tokens
- *CountQU / Sentence:* Ratio of subjective sentences from subjectivity dataset to all sentences
- *Count PL / Sentence:* Ratio of objective sentences from subjectivity dataset to all sentences
- *Average POS Tags / Sentence:* Average number of POS tags per sentence
- *Adjectives / Token:* Ratio of number of adjectives to number of all tokens

- *Adverbs / Token:* Ratio of number of adverbs to number of all tokens

The results for the best 16 features, according to the highest MI, are shown in Figure 5.11. For all further experiments, only the features with the highest MI have been taken into consideration.

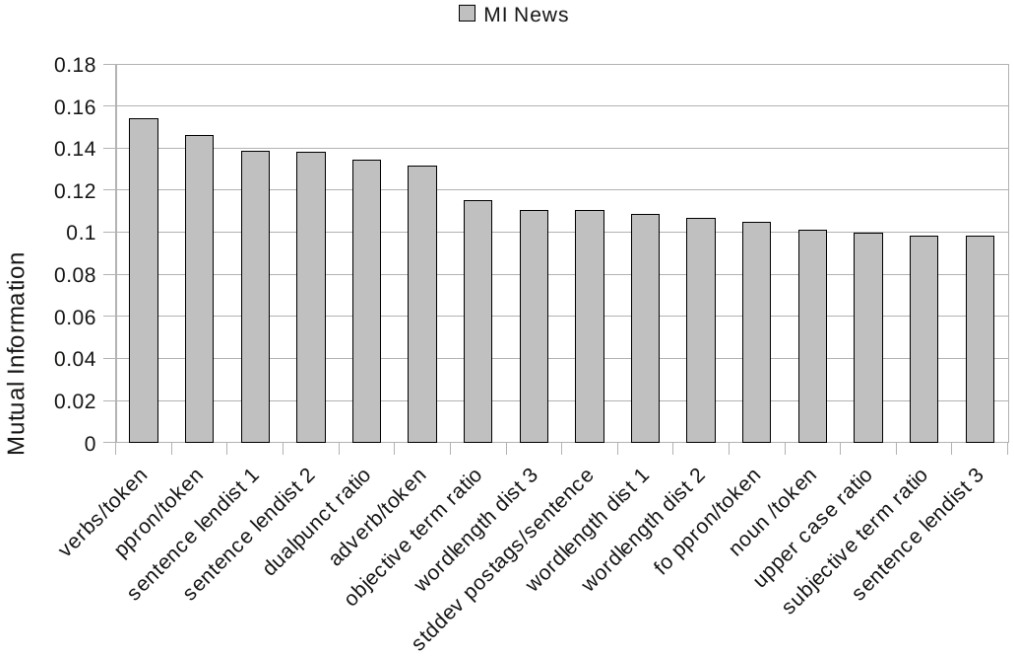


Figure 5.11: Mutual Information of Top16 Stylometric Features

The results for the uni-domain and the cross-domain quality classification setting with stylometric features are shown in Figure 5.12. This experiment has revealed that overall, the accuracy values achieved with stylometric features are lower than with lexical features - both for uni-domain as well as cross-domain. However, in the cross-domain task, the accuracy diminishes only by a few percent in comparison to the single domain task. Consequently, the stylometric feature provide a higher level of generalizability compared to the lexical features - especially, when the topics of training and test set are different.

Lessons Learned

For the quality classification experiment, lexical as well as stylometric features have been evaluated in a uni-domain as well as a cross domain setting. As datasets, online news articles have been used. From the application of both feature types in the

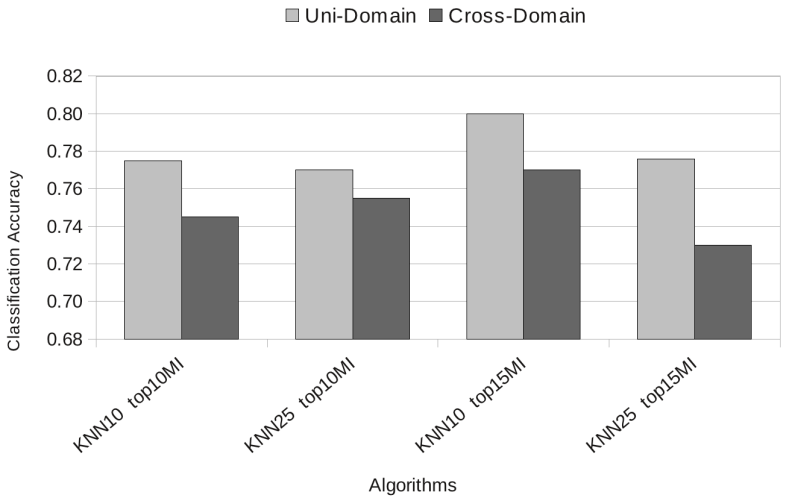


Figure 5.12: Performance on News: Classification Accuracy of KNN with different values for k with Stylometric Features in Uni-Domain and Cross-Domain Setting

uni-domain and the cross-domain settings it has been derived that lexical features implicitly learn topics instead of quality characteristics. On the contrary, the stylometric features are generalizable over different topics albeit with lower classifier accuracy. The experiments showed that stylometric features achieve an accuracy of 77% whereas lexical features reach a nearly perfect classification of 95%. Summing up, if it is guaranteed that topics stay the same over time, lexical features are the solution at hand. If the focus is on being as topic independent as possible, it is advisable to use stylometric features.

5.5.2 Objectivity Classification in Blogs

The same features as in the quality classification in Online news are used to assess the objectivity of news related blogs. According to the formal description of the content facet objectivity in Section 4.2.3, objectivity and quality are typically closely related; therefore, the same features and algorithms should be applicable. Also, stylometric features should be suitable in this context: For instance, the number of first person pronouns is a good indicator for the objectivity of a blog [Lex et al., 2010c]. Consequently, this experiment also investigates whether the proposed features can be used in another domain for a similar task.

Dataset

The blog corpus again consists of a subset of the TREC Blogs08[12]. dataset. Since to the best of the author's knowledge, no standard blog corpus exists for objectivity classification, a randomly chosen subset of the Blogs08 dataset has been annotated that consists of 83 English blogs and 12844 distinct blog entries.

In the annotation step, the blogs have been annotated into the classes *objective* and *subjective*. Note that each blog has been labeled with its main category. In other words, if an overall blogs contained a few subjective blog entries, the whole blog has been labeled as being objective.

To validate the correctness of the labeling step, the corpus has been analyzed in detail. More specifically, its verbs, adjectives and adverbs have been compared with the terms of a gold standard subjectivity lexicon [Wilson et al., 2005]. As a result, it has been found that 40.8% of the verbs, adjectives and adverbs (64279 terms) from the subjectivity lexicon were also present in the subjective blogs. In contrast to that, the objective blogs only contained 18.5% of the verbs, adjectives and adverbs (55805 terms) in the subjectivity lexicon. Since blogs labeled as subjective contains significantly more subjective terms than blogs labeled as objective, this enables to draw the conclusion that the manual annotation of the blog corpus should be correct.

This has been investigated further by comparing a selection of the used features derived from the blog corpus with the subjectivity labeling of the Movie-Reviews corpus [Pang and Lee, 2004]. Movie-Reviews consists of 10k independent sentences, among them 5k objective and 5k subjective sentences.

Figure 5.13 shows the normalized ratio between objective and subjective for different features. Note that the Movie-Review corpus is referred to as Extern in the Figure. This experiment reveals that the ratios are highly correlated. This underpins that the annotations of the blog corpus are applicable for objectivity classification in blogs.

Approach

In this experiment, both lexical features (tokens, unigrams, bigrams) as well as the proposed stylometric features have been evaluated [Lex et al., 2010a]. Additionally, the terms from the earlier described subjectivity lexicon have been used as features

[12] `http://ir.dcs.gla.ac.uk/test_collections/blogs08info.html`

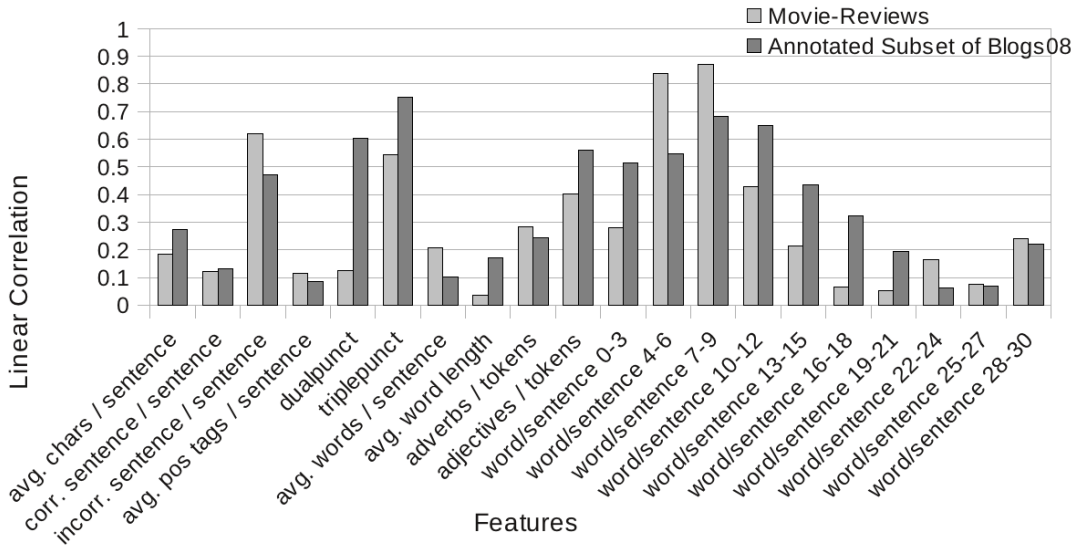


Figure 5.13: Corpus Correlation between the Labels of the Annotated Subset of the Blogs08 Corpus and the Subjectivity Labeling of the Movie-Reviews Corpus

whereas their number of occurrence has been computed relation to the total number of tokens in a document.

The results for the best 16 features, according to the highest MI, are shown in Figure 5.14.

Experiments and Results

Figure 5.15 shows the performance of the best performing algorithms on the blog corpus.

As one can see, the best performance is again achieved on tokens, unigrams, and bigrams with the SVM. However, the classifier model reflects topics, as investigated earlier, and therefore this approach has to be used with care. The CFC performs second best on trigrams due to the highly sparse nature of the trigram vectorspace. Note that all accuracy values were 10-fold cross-validated.

On the stylometric feature set, the achieved classification accuracy is lower, but good with up to 85% accuracy. Note that those features actually reflect characteristics that indicate objectivity and that they are completely topic-independent, as has been shown in the previous quality classification experiment. The k-NN performs best with $k = 25$ on the top 15 features.

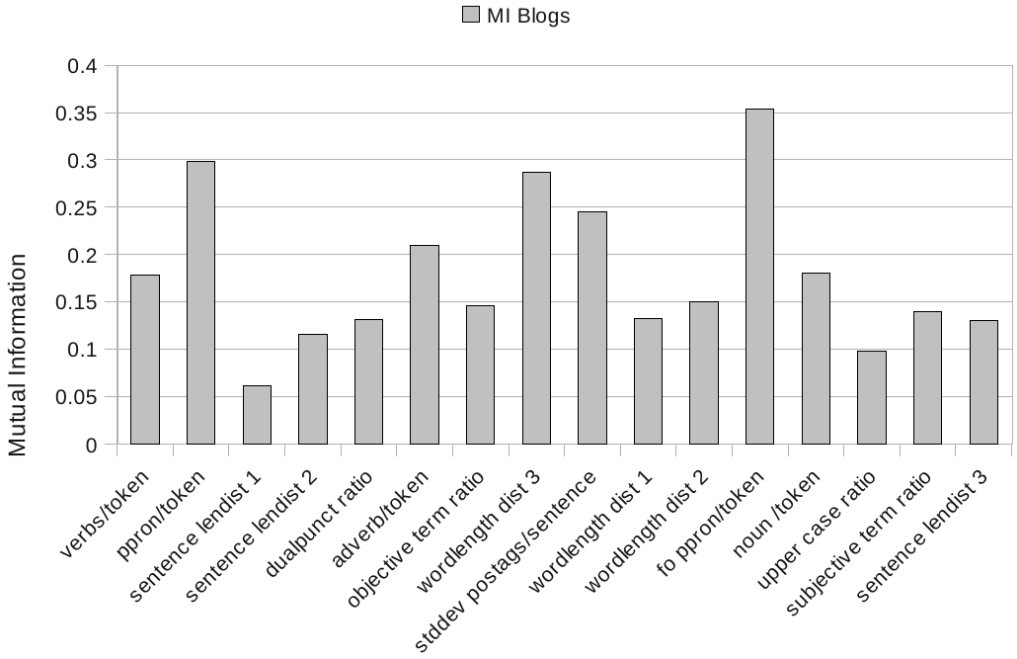


Figure 5.14: Mutual Information of Top16 features

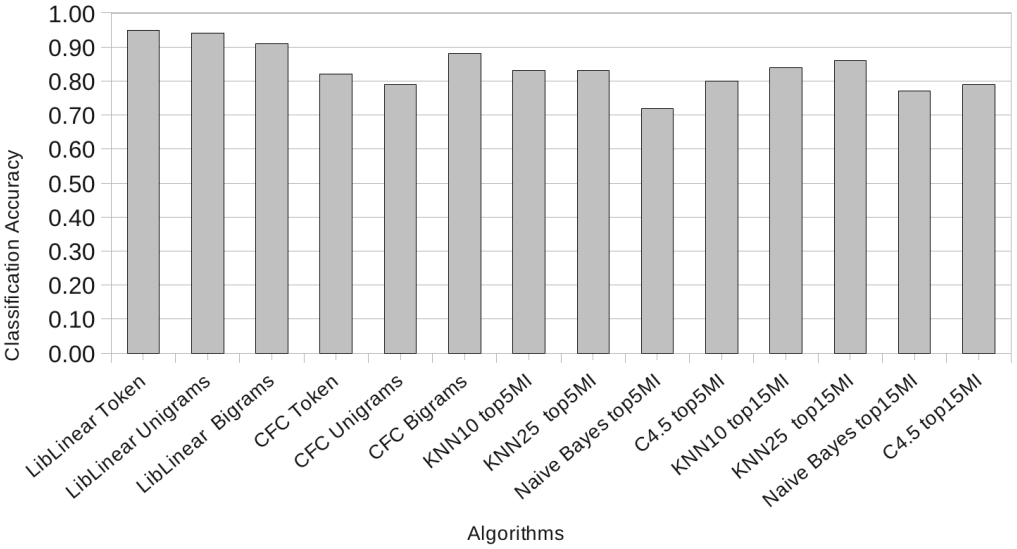


Figure 5.15: Classification accuracy on the blog corpus, achieved with the best performing algorithms

Lessons Learned

Lexical as well as stylometric features used for the earlier described quality classification experiment in Online news have be applied to a similar task; more specifically

to objectivity classification in blogs. As outlined earlier, objectivity is typically a strong indicator for quality. The experiments revealed that the proposed feature and algorithm combinations are applicable also in another domain in a similar setting.

5.5.3 Emotion Classification in Blogs

This section describes the assessment of the topic independent content facet emotionality. More specifically, in the course of this dissertation research, the emotionality of social media, blogs respectively, has been assessed [Lex et al., 2010a, Lex et al., 2010b] following the formal description of the content facet emotionality in Section 4.2.3. Note that the assessed blogs all have been related to news and current events and this experiment is based on the single-domain genre classification experiment outlined in Section 5.4.3.

The emotionality within these news related blogs can be exploited to identify the feelings of individuals toward specific events. Especially for media resonance analysis, it is highly important to determine the reaction towards certain events or campaigns. If an author blogs emotionally, the event definitely concerns her and therefore this event naturally attracts more attention. Besides, media consumers should have the possibility to filter news related blogs by emotionality since in some cases, they may be interested in e.g. emotional eye witness accounts to current events like the 2011 Tsunami in Japan[13] and in other cases, they may want to read factual blog posts about what to do in case of a Tsunami[14].

Dataset

Due to the lack of a standard corpus related to this experiment, the same subset of the TREC Blogs08 Dataset[15] has has been used in Section 5.4.3 for the single-domain genre classification approach. The 83 blogs have been manually annotated with total number of 12844 distinct blog entries in English into *Emotional* versus *Neutral*. Figure 5.7 shows the corpus distribution for this task.

The corpus distribution table reveals that the corpus is fairly balanced for both classes of interest, Emotional and Neutral.

[13] `http://blogs.chron.com/newswatch/2011/03/pacific_tsunami_on_twitter.html`

[14] For an example of such a blog post, see `http://www.jonobacon.org/2011/03/11/japan-tsunami-what-to-do-if-it-affects-you/`, last accessed March 2011

[15] `http://ir.dcs.gla.ac.uk/test_collections/blogs08info.html`

Algorithms

The following text classification algorithms have been applied and evaluated: a Support Vector Machine (SVM) based on LibLinear [Fan et al., 2008], a Support Vector Machine (SVM) based on LibSVM [Chang and Lin, 2001], a k-NN algorithm [Aha et al., 1991] with $k = 10$ and cosine similarity as distance measure in case of lexical features and Euclidean distance in case of stylometric features, and the Class Feature Centroid (CFC) with $b = 1.1$ [Guan et al., 2009]. Additionally, the Mallet implementation of a Naive Bayes classifier, with and without Boosting (AdaBoost) as well as a C45 Decision Tree with and without Boosting (AdaBoost) have been used.

Approach

The approach for this experiment has again been to apply lexical (see Section 5.3.1) versus stylometric text features (see Section 5.3.2) to classify news related blogs into emotional versus neutral. Note that in this thesis, this is further referred to as Emotionality Assessment (EA) task.

In the course of the EA task, the statistical properties of the proposed stylometric features have been analyzed. For this work, stylometric features have been chosen that specifically reflect emotion characteristics. For instance, as stylometric features, the distribution of emoticons over the corpus has been used. Note that emoticons are a series of punctuations and characters which many browsers and tools interpret as different smileys.

Similarly as in the earlier described content facet experiments, it has been investigated which features are most relevant for the EA task. For instance, the number of adverbs per tokens is a good indicator for the emotionality of blogs.

	Emotional	Neutral
blog level	52%	48%
entry level	40%	60%

Table 5.7: Corpus Distributions of the Blog Corpus for the Emotion Assignment Task

Statistical Feature Analysis

To identify the most relevant stylometric features, two established statistical measures have been used; namely the Linear Correlation (LC) and the Mutual Information (MI). The results of this experiment is shown in Figure 5.16. For the EA

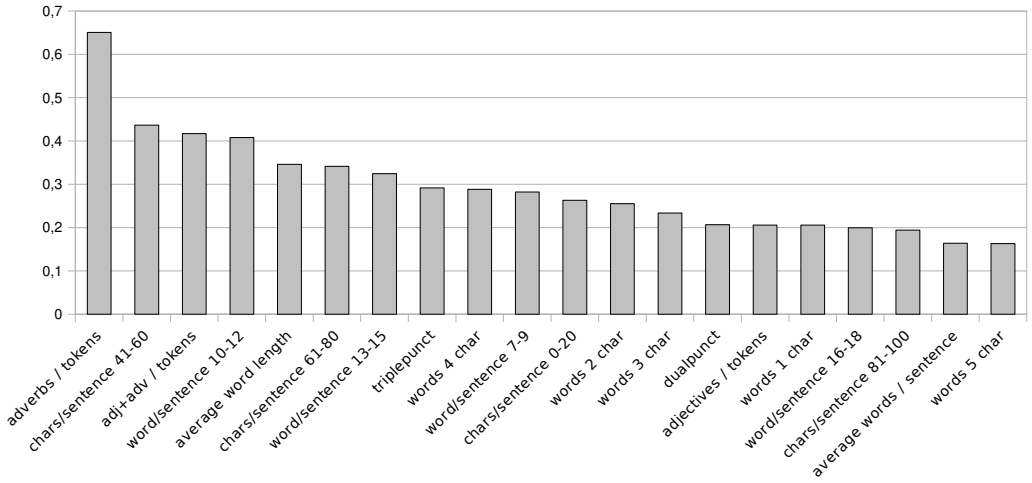


Figure 5.16: EA Task: Linear Correlation

task, the *adverbs/token* feature is most correlated. This is interpreted so that an individual who writes very emotionally often uses many adverbs whereas a rational writer uses adverbs rarely.

The eight best feature is the triple punctuation feature which covers emoticons often used in blogs.

As second statistical measure, the Mutual Information (MI) has been used. Figure 5.17 shows the mutual information for the EA classification task.

Apparently, the MI reveals that the number of used adverbs is also the most correlated feature. In this case, the *lower case/upper case* feature is the second most correlated feature, and features denoting the word/sentence complexity (*Characters per Sentences* and *Words per Sentences*) are also highly correlated.

The next section describes the results that have been achieved using the different features and algorithms.

Classification Results

For the EA task, at first the lexical features have been evaluated: both on blog level as well as on blog post level. The results achieved on blog level are shown in

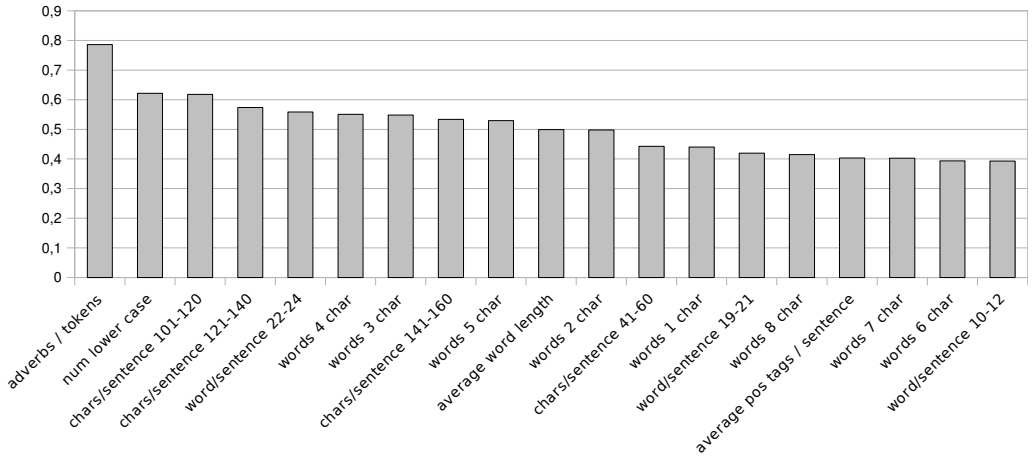


Figure 5.17: EA Task: Mutual Information

Figure 5.18 in terms of classification accuracy.

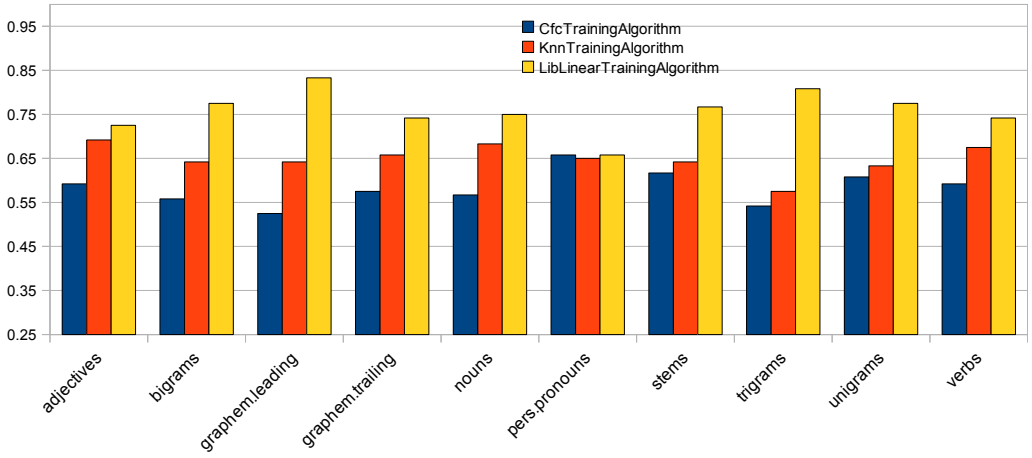


Figure 5.18: Emotion Classification in Blogs: Classification Accuracy with Lexical Features on Blog Level

Figure 5.19 depicts the results achieved with lexical features on blog post level. These two experiments confirm the observations from the single-domain genre classification (see Section 5.4.3: Clearly, there is a difference on which level the experiments are conducted (blog level LibLinear on graphems 83.0 versus blog post level LibLinear on stems 91.4).

To compare the performance of stylometric features and lexical features, the stylometric features have also been evaluated, using the same classification algorithms as well as others that are especially suited for the dense stylometric feature spaces.

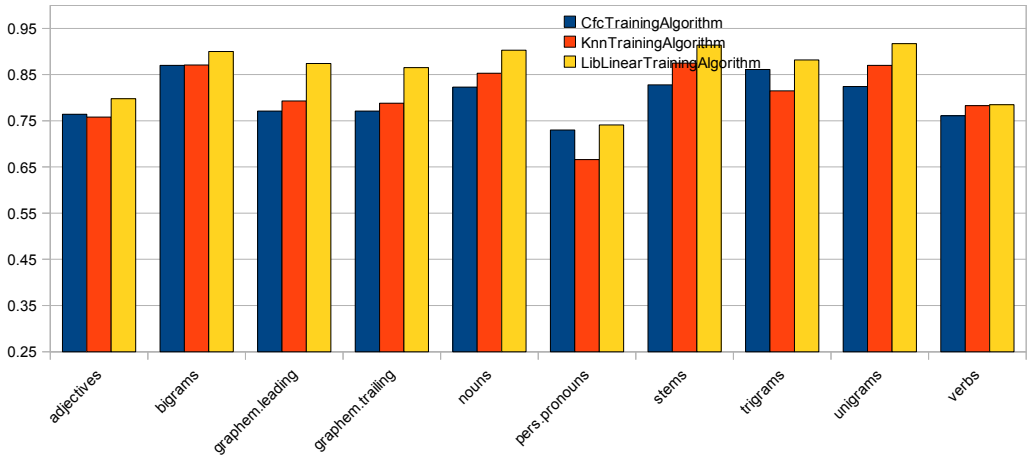


Figure 5.19: Emotion Classification in Blogs: Classification Accuracy with Lexical Features on Blog Post Level

More specifically, the Mallet implementation of a Naive Bayes classifier, with and without boosting (AdaBoost) as well as a C45 decision tree with and without boosting (AdaBoost) have been used. Due to the fact that a linear kernel is often not the optimum for dense data, also applied LibSVM [Chang and Lin, 2001] has been applied on the classification problem with the more general RBF kernel, as suggested by the LibSVM user guide. Further, a grid search for the SVM parameters C ($2^{-3} - 2^5$) and γ ($2^{-6} - 2^3$) has been performed. The best performing parameter set is $C = 2^5$ and $\gamma = 2^1$.

The results on blog post level for the different classifiers with lexical features are shown in Figure 5.20. For these experiments, only the features with the highest Mutual Information and Linear Correlation have been taken into account. This resulted in a number of distinct features of 24.

The experiments with the different classifiers corroborated earlier findings: with stylometric features, a lower classification accuracy is achieved for both tasks compared to the high dimensional feature space classifiers based on lexical features. Nevertheless, stylometric features are guaranteed to be topic independent and therefore their validity is higher in case a topic independent content facet like emotionality is assessed [Lex et al., 2010c].

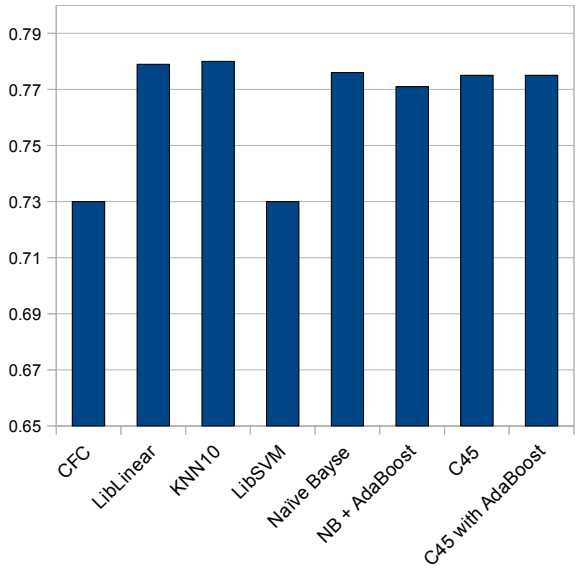


Figure 5.20: Emotion Classification in Blogs: Classification Accuracy with Stylometric Features on Blog Post Level

Lessons Learned

In this experiment, the content facet *emotionality* has been assessed. For this, the performance of lexical and stylometric features has been evaluated in news related blogs. The impact of this experiment is: Firstly, the emotionality of news related blogs has to be assessed on blog entry level. Secondly, topic independent stylometric text features can be used to perform this task albeit with a lower accuracy than achieved with lexical feature but with the advantage of being guaranteed topic-independent.

5.5.4 Blog Credibility Ranking in News

"Don't know what credibility is but know what it feels like to lose it"[16]

In this experiment, the topic independent content facet credibility has been assessed in news related blogs [Juffinger et al., 2009a] following the formal description of the content facet credibility in Section 4.2.3.

The general approach for this experiment is to compare two statistical properties, namely the quantity structure and the content of blogs with a credible source. In

[16]Douglas Feaver, Executive Editor, Washington Post

this context, the quantity structure denotes the time series of a blog, that is the number of blog post that have been published to a given topic at certain time units. Exploiting the quantity structure in the context of blog credibility assessment has been introduced as novel dimension within this dissertation research. The rationale behind this dimension is that the quantity structure enables to filter out blogs with a totally different publishing behavior than a credible source. Such blogs can be for example spam blogs or advertising blogs.

For this experiment, a quality checked and verified reference news corpus from the Austrian Press Agency (APA) has been used as credible source.

Based on the proposed methodology, blogs have been ranked by credibility into three levels: (i) highly credible, (ii) average credible, and (iii) little credible.

Dataset

The verified news corpus consists of German news articles from the news article repository provided by the Austrian Press Agency (APA) and it contains APA news articles that have been downloaded within a time period of 2 months in 2008.

The blog corpus consists of 40 blogs that have been crawled from the Web in the same time period as the APA corpus. Special attention has been paid to use blogs that are related to news and current events. For this experiment, the used blogs have been selected manually in order to guarantee that they are indeed related to news and current events.

Credibility Ranking Process

The blog credibility ranking procedure starts with a search query in the German news corpus as well as in the blog corpus. As a result, German news articles as well as blogs are retrieved.

The search query terms have been manually selected whereas for instance common politician names have been used. The following query terms have been used:

> Obama, Bush, Sarkozy, Brown, Putin, Haider, Pröll, Faymann, Strache, Glawischnig, Barroso, Medwedew, Hasan, Merkel, Mugabe, Mumbai, Irak, Frankreich, Europäische Union, EU, Athen, Deutschland, Georgien, New York, Paris, Wien, Tschechien, USA, Indien, Tschad, Finanz, Poli-

tik, Krise, Wirtschaft, Bank, Tourismus, Schifahren, Transit, Tarif, Umwelt, Immobilien, Wahl, Unfall, Terrorismus, Gesundheit

The retrieved German news articles are then used as credible basis. Since the APA articles are only available in German, the credibility ranking in this experiment has also been restricted to German blogs.

Both the news articles as well as the blogs are then subject to a two-stage process. Firstly, a quantity structure analysis and filtering step based on a timewarping procedure and correlation coefficient calculation is performed (see Section5.5.4). Secondly, a content similarity evaluation provides a credibility ranking (see Section 5.5.4). The overall process of blog credibility ranking is shown in Figure 5.21. The time warping procedure is outlined in more detail in the next paragraph.

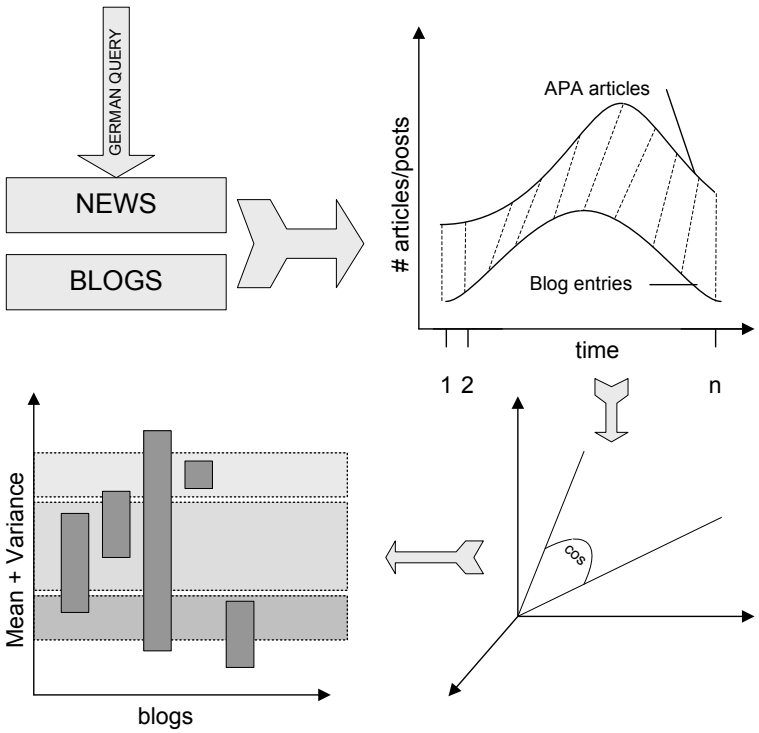


Figure 5.21: General Process of Blog Credibility Ranking.

Time Warped Correlation In order to make the news corpus and the blog corpus comparable, both corpora have to be aligned first. This has two reasons: Firstly, they typically differ in the amount of posts or articles per time. Secondly,

they are shifted in time since sometimes an event is first published in traditional media and later discussed in social media or the other way round (see Figure 5.21).

The alignment of the corpora is based on the Dynamic Time Warping (DTW) algorithm. The DTW can be used to compute whether two time series exhibit similarities or to identify corresponding regions of two time series. Due to performance reasons, a FastDTW algorithm has been applied that is based on the work of Fu et al. [Wai-chee et al., 2005] and [Sakoe and Chiba, 1978].

Given two time series, N and B of lengths $|N|$ and $|B|$, a Warp path has to be constructed so that

$$W = w_1, \ldots, w_K \tag{5.2}$$

whereas $max(N, B) \leq K < B + N$; K denotes the length of the Warp path W. The k_{th} element $w_k = (i, j)$ where i denotes the index from time series N and j the index from the time series B.

The goal of the DTW algorithm is to identify the optimum warp path defined by the minimum distance warp path given here:

$$min(Distance(W)) = min(\sum_{k=1}^{k=K} Distance(w_{ki}, w_{kj})) \tag{5.3}$$

Note that the distance $Distance(W)$ is typically derived by computing the Euclidean Distance of the Warp path W. The expression $Distance(w_{ki}, wkj)$ gives the distance between two data point indices from N and B in the k_{th} element of the warp path [Salvadore and Chan, 2004].

Since the DTW algorithm is quadratic in time and space complexity, the Fast Dynamic Time Warping algorithm has been introduced in [Salvadore and Chan, 2004]. The FastDTW introduces constraints and means to perform data abstraction so that the DTW is computed on a representative subset of the original data. Besides, bounding functions have been introduced in order to reduce the number of runs for the DTW. The FastDTW algorithm takes as input two time series N and B of length $|N|$ and $|B|$. Besides, the parameter *radius* defines the distance within the warp path is refined by the algorithm. The output of the FastDTW algorithm is a warp path with minimum distance between the time series N and B [Salvadore and Chan, 2004].

As showcase for the validity of the time warping serve Figure 5.23 and 5.22. In these two examples, the temporal distribution of news articles and blog posts over time. From the figures, it can be derived that both corpora apparently are correlated in respect to their quantity structure.

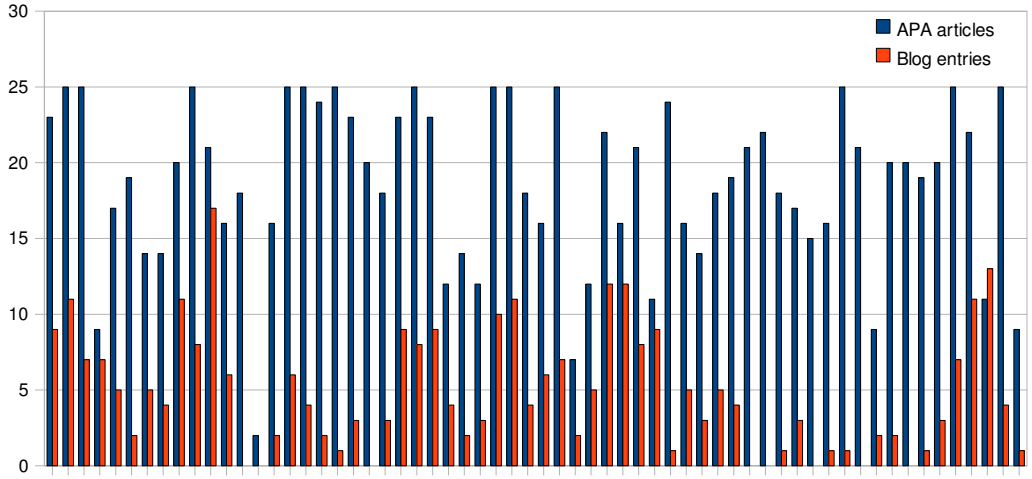


Figure 5.22: Temporal Distribution of the News Articles and the Blog Posts for Query "Frankreich".

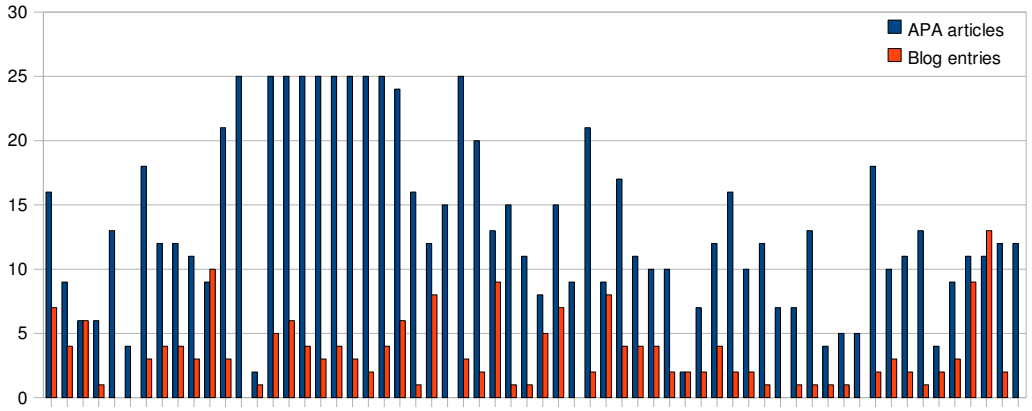


Figure 5.23: Temporal Distribution of the News Articles and the Blog Posts for Query "Obama".

In these examples, the quality of the selected blogs has been rather high. An interpretation of this is that a strong correlation of the quantity structures can only be found when the selected blogs do not consist of advertisement or spam blogs but of quality news-related items.

In order to filter out blogs that exhibit a totally different quantity structure, the Pearson's product moment correlation coefficient[17] has been computed between news and blogs. Based on this correlation coefficient, blogs are filtered out that exhibit a significantly different quantity structure for a certain query. The determination of the correlation between news and blogs is described in Algorithm 2.

Data: The set R of search results r_i retrieved from blogs and news corpora for the queries q_a of Q. Each retrieved set $R_{1a} \subset R$ and $R_{2a} \subset R$ is represented as time series and ordered in ascending time. Multiple results per time entry are possible.
Result: The aligned correlation between R_{1a} and R_{2a}.

```
foreach q_a in Q do
    foreach r_i in R_1a do
        newsResults[i] = #QueryResults(r_i) ;
    end
    foreach r_i in R_2a do
        blogResults[i] = #QueryResults(r_i) ;
    end
    FastDTW(newsResults, blogResults, window);
    correlation[a] = pearsonsCorrellation (newsResults,blogResults);
end
```

Algorithm 2: Query Results Alignment

Especially the time warping step is crucial; for example, the correlation between news and blogs to the query Frankreich, shown in Figure 5.22, results in a correlation coefficient of 0.23 if no time warping is performed as opposed to a coefficient of 0.79 if time warping is performed beforehand.

Based on a Leave-One-Out (LOO) strategy, the correlation coefficient for the news corpus has been compared with the correlation coefficient of each blog. Whenever a significantly higher coefficient has been achieved without a blog, this blog has been removed. The robustness of the correlation coefficient against a constant offset is thereby helpful, because possibly credible blogs with a constant amount of entries per day do not have an impact on the correlation coefficient and are therefore not sorted out. Blogs with actual events shortly after or before the news papers remain also, due to the earlier described time warping step. Note that only blogs with a

[17]Wolfram Mathworld, Correlation Coefficient, `http://mathworld.wolfram.com/CorrelationCoefficient.html`, last accessed May 16, 2011

negative influence on the correlation have been removed since they are the blogs with a completely different distribution over time.

Figure 5.24 and Figure 5.25 show the number of blog entries and news articles according to 30 different queries in a specific time period. The following queries have been used:

> Obama, Bush, Sarkozy, Brown, Putin, Haider, Pröll, Faymann, Strache, Glawischnig, Barroso, Medwedew, Hasan, Merkel, Mugabe, Mumbai, Irak, Frankreich, Europäische Union, EU, Athen, Deutschland, Georgien, New York, Paris, Wien, Tschechien, USA, Indien, Tschad

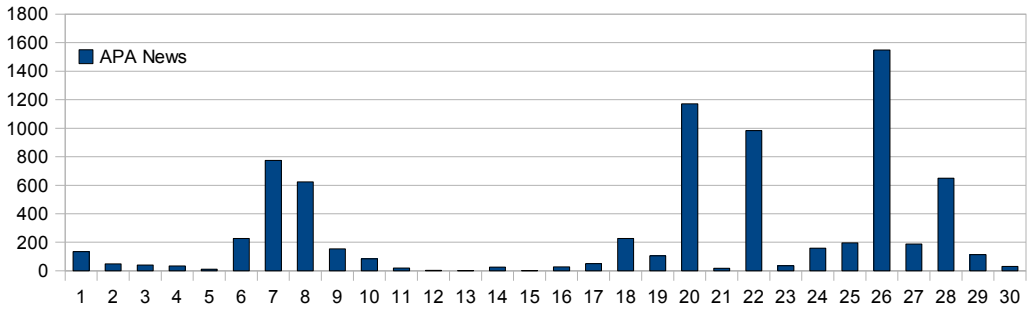


Figure 5.24: Quantity Structure of different Queries in APA Articles.

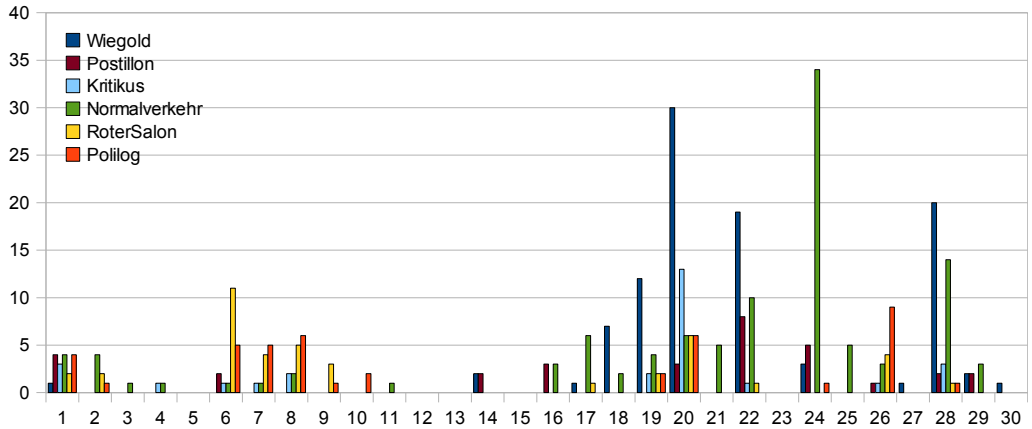


Figure 5.25: Quantity Structure of different Queries in Blogs.

Table 5.8 provides the numeric values of the correlations between the news corpus and particular blogs. From Figure 5.24 and Figure 5.25 it can be derived there is a correlation between the quantity structure of news and selected blogs. Table 5.8

Blog	Original	Warped
Wiegold	0.547	0.476
Polilog	0.750	0.507
Roter Salon	0.491	0.301
Normalverkehr	0.180	0.462
Postillon	0.376	0.214
Kritikus	0.554	0.171

Table 5.8: Results: Content Correlation for 30 Queries between selected Blogs and the News corpus

reveals that especially the blogs *Polilog*, *Wiegold*, and *Normalverkehr* are highly correlated with the verified news corpus.

After the time warped correlation filter, a set of blogs with appropriate quantity structure has been derived. The time warped correlation filter does not take into consideration any content wise information. Therefore, this resulting set of blogs has further been subject to an in-depth content wise analysis. This is described in the next paragraph.

Content Correlation To perform an in-depth content wise analysis, this thesis proposes to measure the content similarity between credible news and blogs. For this, both the news articles as well as the blogs have been preprocessed using techniques from Natural Language Processing (NLP). The preprocessing is described in the next paragraph.

Content Preprocessing Firstly, the news articles and the blog posts have been tokenized, Part-Of-Speech (POS) tagged, and stemmed. As features, the POS nouns, adjectives, and verbs have been used. The rationale behind that has been that nouns generally cover the topic wise information of a document while verbs and adjectives indicate the association with the topic.

The result of this content preprocessing step is then a list of stemmed terms per document with the absolute term frequency (ATF) of each term. For each document, a term vector has been created whereas the vectors have been weighted using TF-IDF and normalized to unit length. Then, for all news articles, the centroid vector has been computed. For each blog, also a centroid vector [Han and Karypis, 2000b] has been derived whereas it has been summed over the blog entries of the blog.

The centroids are then again normalized to unit length and then, the cosine similarity in-between the blog centroids and the news centroid has been computed. This is outlined in the next paragraph.

Centroid Cosine Similarity Generally, the cosine similarity is an established means to derive the similarity between two vectors in a high dimensional vector space [Qiu and Pang, 2008]. The computation of the cosine similarity is shown in Equation 5.4 whereas d denotes the vector space representation of the documents [Han and Karypis, 2000b].

$$similarity = cos(d_i, d_j) = \frac{d_i.d_j}{\|d_i\|_2\|d_j\|_2} \tag{5.4}$$

In the proposed methodology to rank blogs by credibility, for each search query, a set of blogs and news articles has been derived. In addition to the computation of the similarity between the particular document sets, the mean and the variance has been derived over the set of search queries. Note that the queries have been categorized into groups of topics: for instance, to derive the credibility of political blogs, the search queries denoted names of politicians, political regions, and political concepts.

The content similarity results have been further been used to assign one out of three proposed credibility levels to each blog. The blog credibility ranking methodology is described in the next section.

Blog Credibility Ranking Methodology

As mentioned earlier, the content wise similarity computation has been based on different POS; namely (i) nouns, and (ii) verbs plus adjectives.

For the blog credibility assessment, in the first step, only the similarity values derived based on nouns have been taken into account. All mean and variances have been compared whereas a threshold has been applied to sort out blogs with the lowest similarity values.

The threshold is determined by the interval defined in Equation 5.5.

$$t = \left[0, min(sim_j) + \frac{1}{2} * (max(sim_i) - min(sim_j))\right] \tag{5.5}$$

All blogs below the threshold are assigned to the credibility level *little credible*. Then, on the remaining blogs, the same procedure is applied to the similarity values based on verbs and adjectives, also to assign blog to the credibility level *little credible*. On the remaining blogs, again, the content correlation step is computed whereas in this case without restrictions on any POS.

Note that a manual examination of the resulting little credible blogs revealed that these blogs either deal with different topics (from nouns) or are in a completely different association with the topic (from verbs and adjectives).

All blogs within the interval given in Equation 5.6 have been assigned to the credibility level *highly credible*. As an additional constraint, the maximum variance of highly credible blogs has been set to be less than 0.05.

$$t = \left[min(sim_j) + \frac{1}{2} * \left(max(sim_i) - min(sim_j) \right), 1 \right] \tag{5.6}$$

Finally, all remaining blogs that neither fall into the credibility level *little credible* nor *highly credible* have been assigned to the credibility level *average credible*.

Experiments and Results

In the context of the the blog credibility ranking, two scenarios have been evaluated. These scenarios are described in the next paragraph.

Scenarios For the news corpus, all news articles that resulted from the particular search query have been considered as being relevant. In case of the blog corpus, however, not all posts of a blog might be relevant to a search query. More specifically, while a blog can match a search query, its blog posts might contain items that deal with another topic than the search query. Since the goal of this experiment has been to rank the whole blog by credibility, this issue has been investigated further.

For this, two settings have been evaluated that use either all blog posts or only blog posts that are relevant to the search query:

- **Setting1:** Only the blog posts that are actually relevant to the search query are used to rank the blog.
- **Setting2:** All blog posts of a blog are used to rank the blog.

These two settings have been evaluated based on the Cosine Similarity between the particular news and blogs. The results for Setting1 are shown in Figure 5.26. In Figure 5.27, the content Cosine Similarity mean and variance values for Setting2

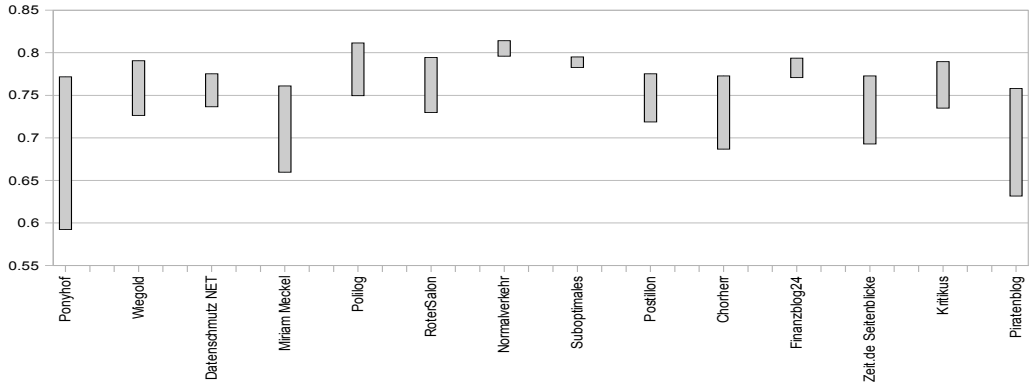


Figure 5.26: Blog Credibility Ranking, Setting1: Cosine Similarity Mean and Variance for Relevant Blogs.

are shown. A comparison of both figure reveals that a credibility ranking from the

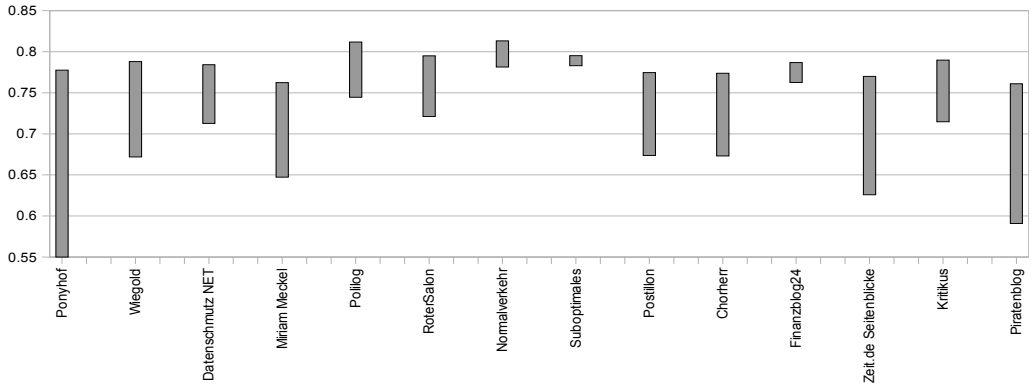


Figure 5.27: Blog Credibility Ranking, Setting2: Cosine Similarity Mean and Variance for Relevant Blog Posts.

similarity values between the news corpus and only the relevant blog posts (Setting1) is not feasible. In other words, if one blog post perfectly matches the news corpus, the full blog not necessarily does. Therefore, for the further experiments, only Setting2 has been used.

In the next experiment, the content correlation procedure is computed based on nouns only. Figure 5.28 shows the results for this experiment. The blogs below the threshold given in Equation 5.5 can be clearly identified. As a visual support, in the

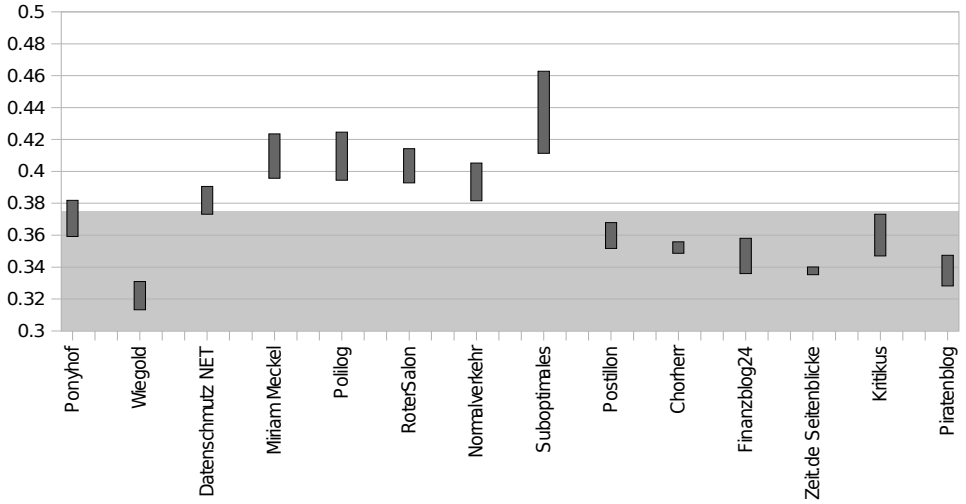


Figure 5.28: Blog Credibility Ranking: Centroid Cosine Similarity Mean and Variance for Nouns.

figure this area is highlighted in gray. More specifically, these blogs are Ponyhof, Wiegold, Postillion, Chorherr, Finanzblog, Seitenblicke, Kritikus and Piratenblog. Also, it can be seen that all blogs except Ponyhof clearly are below the threshold. The small variance of these blogs also indicate that the blogs consistently deal with similar topics.

In the next experiment, the content correlation procedure is computed based on verbs and adjectives only. The calculation of the threshold based on these features leads to nearly no additional assignments. This is due to the correlation between nouns, verbs and adjectives in natural language texts. In other words, even if documents are off-topic, they nevertheless exhibit a small similarity in respect to nouns, adjectives, and verbs. Therefore, in the next step, firstly, the content correlation is computed based on nouns and the blogs are labeled with the credibility levels. Secondly, on blogs that could not be assigned to a credibility level, the content correlation based on verbs and adjectives has been computed. In other words, in a first step, a noun filter has been applied and in a second step, an adjectives plus verb filter.

Figure 5.30 shows the results of the content correlation based on verbs plus adjectives on the remaining blog that could not be assigned with the noun filter. Again, the threshold has been applied and this results in an assignment of another

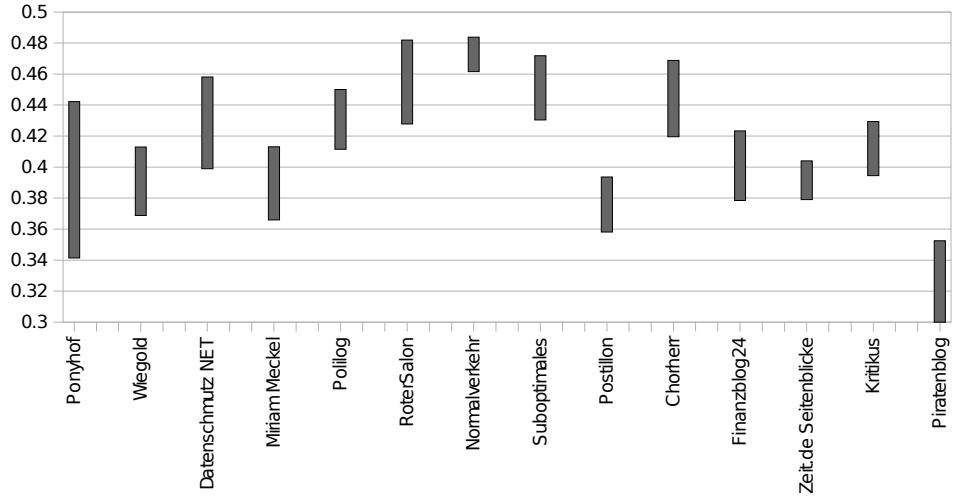


Figure 5.29: Blog Credibility Ranking: Centroid Cosine Similarity Mean and Variance for Adjectives and Adverbs.

blog, the Miriam Meckel blog to the credibility level *little credible*.

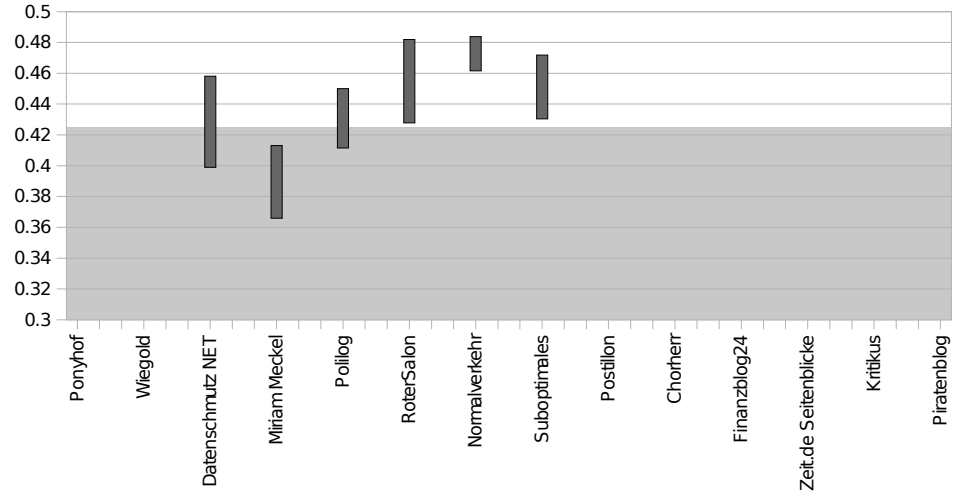


Figure 5.30: Blog Credibility Ranking: Centroid Cosine Similarity Mean and Variance with Adjectives and Adverbs Filter on Blogs that could not be assigned to the Credibility Levels with the Noun Filter.

The blogs above the threshold and with low variance are classified as *highly credible*: Polilog, Roter Salon, Normalverkehr, and Suboptimales. According to the proposed methodology, the remaining blogs fall into the middle. Consequently, they

have been assigned to the credibility level *average credible*. The results of the blog credibility ranking process is also summarized in Table 5.9.

Blog	Highly Credible	Average Credible	Little Credible
Ponyhof			X
Wiegold			X
Datenschmutz NET		X	
Miriam Meckel			X
Polilog	X		
RoterSalon			X
Normalverkehr	X		
Suboptimales	X		
Postillon			X
Chorherr			X
Finanzblog24			X
Zeit.de Seitenblicke			X
Kritikus			X
Piratenblog			X

Table 5.9: Results: Blog Credibility Ranking

To verify the results of the blog credibility ranking, domain experts ranked the investigated blogs into the three credibility levels. All nine blogs assigned to the credibility level *little credible* by the proposed approach have also labeled as *little credible* by the domain experts. Unfortunately, the proposed methodology ranked too little credible blogs incorrectly, namely Polilog and Roter Salon.

These two falsely ranked blogs have been investigated in more detail. This investigation revealed that the structure as well as the wording of both blogs has been quite similar to credible news articles. Consequently, they have been falsely assigned even though their content is not credible at all. To assign a correct credibility value to these remaining blogs, it is therefore necessary to incorporate other content facets as well, e.g. quality.

To sum up, the proposed blog credibility ranking methodology has enabled to filter out 26 inappropriate blogs from 40 blogs solely based on their publishing behavior over time denoted by the quantity structure. From the remaining 14 blogs, 12 blogs have been assigned to one of the three proposed credibility levels, in accordance with the judgement of the domain experts. Therefore, the precision for the credibility level *little credible* is therefore 0.5, for *average credible* it is 1.0 and for

highly credible it is also 1.0. Consequently, this leads to an average precision of 0.83 for the blog credibility ranking.

Lessons Learned

In conclusion, the proposed blog credibility ranking system has enabled to automatically rank blogs into three levels of credibility. The blog credibility has been determined by exploiting the quantity structure and the content similarity in reference to a credible German news corpus.

The evaluation results have indicated a high quality assignment to the three credibility levels with an average precision of 0.83 on 14 blogs. Therefore, for certain news-related blogs, the assessment of the content facet credibility has been successful for a selection of news related German blogs.

5.5.5 Quality Assessment in Blogs: Combining Multiple Content Facets

In this experiment, the content facet *quality* has been assessed in arbitrary blogs whereas quality in this case is determined by a variety of content facets. This experiment has been conducted in line with the popular Text Retrieval Conference (TREC); more specifically in the context of the TREC Blog Distillation Task 2009 [Lex et al., 2009a]. For details on the TREC conference in general, see Section 2.2.1.

The goal of the Blog Distillation Task 2009 has been to both retrieve blogs that are relevant to predefined topics as well to assess the quality of the retrieved blogs. According to TREC, quality in this context is defined by the following content facets:

- Opinionated: Some bloggers express opinions while other concentrate on factual information. For this facet, the inclinations are *opinionated* versus *factual* blogs.
- Personal: Personal blogs are written in personal time without commercial influences. For this facet, the inclinations are *personal* versus *official*.
- In-depth: In-depth blogs contain in-depth thoughts and a deep analysis of the implications of a topic. For this facet, the inclinations are *in-depth* versus *shallow*.

The goal of the challenge has consequently been to assess the above content facets on a per-topic basis.

Collection and Approach

The experiments for the blog distillation task have been carried out on the Blogs08 dataset. The Blogs08 dataset samples the blogosphere from January 2008 to February 2009. It consists of 28.488.767 blog posts from 1.303.520 English blog feeds [18].

In addition to the Blogs08 dataset, the TREC organizers also provided 50 topics of interest whereas each topic has been assigned to one of the earlier described quality facets. Table 5.10 shows the number of topics for the quality facets. The

Facet	Number of Topics Assigned
Opinionated	21
Personal	10
Indepth	19

Table 5.10: Quality Facets and their Number of Topics assigned.

goal of the TREC Blog Distillation Task 2009 has been to derive three rankings of the most relevant 100 blogs for each of the 50 topics. For the first ranking, the first value of the value was enabled (opinionated, in-depth, personal) whereas for the second ranking, the opposite value of the facet has been enabled (factual, shallow, official). The third ranking served as baseline for the topic relevance which means that only topic relevance and no facets had to be assessed in this case.

The proposed approach has consisted of two steps:

1. Step 1: The first step has been to retrieve blogs that are relevant to a topic or query, respectively. This is actually an Information Retrieval (IR) process.

2. Step 2: The second step has been to assign particular facets of interest to the retrieved blogs.

In Step 1, the retrieval task, 606.939 different blogs out of 1.303.520 blogs have been vectorized and indexed using the open source retrieval framework Apache Lucene[19]. Lucene employs concepts of the Vector Space Model (see Section 2.2.3

[18] http://ir.dcs.gla.ac.uk/test_collections/blogs08info.html

[19] Apache Lucene, `http://lucene.apache.org`, last accessed May 2011

for details) and the Boolean model (see Section 2.2.3 for details) in order to derive whether a document is relevant to a query. More specifically, first the Boolean model is used to derive the subset of documents that contain the query term. On this subset, the Vector Space Model is then used to derive the actual relevancy of a document to the query. The number of occurrence of the query term in a document in relation to the number of occurrence of the query term in all documents in the subset thereby results in a relevance score[20]. For this experiment, this relevance score, the Lucene score, has been used to rank the blogs that have been retrieved to the topics.

From the indexed blogs, 3.476 million different blog feeds have been retrieved. This resulted in a search index size of 41 GB, and the size of the whole repository with vectors for all features has been 220 GB. Since earlier experiments have shown that an analysis has to be performed on blog post level, the single blog posts have been extracted from the blogs by exploiting their permalinks.

With Lucene, the blog index has been searched for the 50 pre-defined topics. Then, the search results have been ranked and the top 100 blogs have been sorted according to the used relevance ranking scheme (see Equation 5.8).

In Step 2, the classification task, firstly a training set had to be created. This training set consisted of a subset of blogs that were randomly taken from the TREC Blogs08 dataset[21] where this set has been manually labeled into the given facets. The classifier was then used to categorize the blogs into the facets. Since the classification has been performed on blog post level, also the annotation has been done on post level. More specifically, if the whole blog was assigned e.g. the facet *opinionated*, this label has been used for all its blog posts. Note that the final classification decision for the whole blog has been achieved using a majority vote over the classifier decisions for the blog posts.

The annotation step resulted in 12844 annotated blog posts. Note that since the labeling of all 12844 blog posts has been checked on only a random selection, it is assumed that there is a certain amount of mis-labeled data.

The unstructured blog posts have then be transformed in a structured form using an Information Extraction and vectorization module based on OpenNLP[22]. In order

[20]Lucene Java Documentation, `http://lucene.apache.org/java/2_3_2/scoring.pdf`, last accessed May 2011

[21]http://ir.dcs.gla.ac.uk/test_collections/blogs08info.html

[22]http://opennlp.sourceforge.net/

to retrieve only English blogs, first the blogs had to be filtered by language whereas a language guesser based on n-grams and the Apache Nutch project[23] has been used.

From the English blog posts, several lexical as well as stylometric features have been extracted. More specifically, nouns, sentences, and punctuation features. For each set of features, a feature vector has been created for each blog post.

Finally, the used classification algorithm has been trained on the annotated blog posts. The learned model has then be used to classify the blogs on blog post level. More specifically, the the top 100 blogs have been classified into the given facets using the trained classifier. As a classification algorithm, a Support Vector Machine based on LibLinear [Fan et al., 2008] has been used with standard parameterization. To assign the desired content facets to the whole blogs, a majority voting over the posts has been computed.

Relevance Ranking

As mentioned earlier, the goal of the TREC Blog Distillation Task has been to rank candidate blogs that are relevant to a topic or query and to assign the retrieved blogs to a set of facets that represent different quality aspects[24].

In order to derive a final relevance ranking, the relevance ranking score derived from Apache Lucene has first been normalized so that its range is between 0 and 1. For this, the Lucene score has been divided by the maximum score which is shown in Equation 5.7.

$$luceneScore = score/maxScore \tag{5.7}$$

Then, the final relevance score has been computed whereas both the Lucene score as well as the classifier confidence, denoted as *facetConfidence*, for the particular facet has been taken into consideration. The formula for the final relevance score is shown in Equation 5.8.

$$finalScore = \alpha * lucenceScore + (1 - \alpha) * facetConfidence \tag{5.8}$$

This final relevance ranking score combines both the Lucene relevance ranking as well as the confidence with which a blog has been assigned to a quality facet.

[23] http://lucene.apache.org/nutch/

[24] http://ir.dcs.gla.ac.uk/wiki/TREC-BLOG

This ranking therefore should favor relevant blogs for which also strong classifier decisions could be derived.

Runs

In the course of the TREC challenge, three runs have been submitted whereas each run is characterized by the features used in this run:

1. *Nounfull NF*: For this run, a feature space based on lexical features has been used, nouns annotated by the OpenNLP library, respectively.

2. *Punctfull PF*: In this run, stylometric features have been used; more specifically a feature space only with punctuation features, e.g. the amount of double or triple punctuations.

3. *Sentencefull SF*: For this run, also stylometric features have been used. A feature space with stylometric properties based on sentence statistics has been used, e.g. the average number of words per sentence, the average number of unique POS tags per sentence, the ratio of lower case characters to upper case characters, the number of adjectives or adverbs per tokens.

Results

The following results have been achieved in the context of the challenge: Out of the 50 topics, 39 topic have been successfully retrieved that had at least one relevant blog for each side of the facet (e.g., one relevant opinionated blog and one relevant factual blog). Consequently, the reported results relate to these 39 topics.

Note that not the whole Blogs08 dataset had been vectorized and indexed[25]. Therefore, relevant blogs were missing in the search index which is reflected in poor results for certain topics.

In Table 5.11, the results for all runs are summarized. For each run, three rankings of the blogs have been performed. The first ranking, denoted as *none*, corresponds to a ranking with no facet value applied and therefore serves as baseline. The second ranking, denoted as *first*, corresponds to the first value of a facet and the third ranking, denoted as *second*, represent the second value of a facet.

[25]Unfortunately, the Lucene index broke a few days before the deadline and the data had to be

RUN	Facet	MAP	R-Precision	P@10
NF	none	0.0624	0.0980	0.1410
NF	first	0.0538	0.0725	0.0769
NF	second	0.0231	0.0346	0.0256
PF	none	0.0624	0.0980	0.1410
PF	first	0.0691	0.0846	0.0692
PF	second	0.0227	0.0339	0.0308
SF	none	0.0624	0.0980	0.1410
SF	first	0.0559	0.0717	0.0667
SF	second	0.0302	0.0334	0.0436

Table 5.11: TREC 2009 Blog Distillation Task Results for All Facets for All Runs

From Table 5.11 it can be derived that the obtained MAP values for all rankings are quite low. In general, MAP corresponds to the average precision and it emphasizes that more relevant documents should be retrieved earlier. Due to the fact that only less than half of the data has been indexed, the retrieval definitely has been suboptimal. More specifically, in many cases, not enough relevant blogs have been retrieved for a topic and this naturally leads to quite low MAP values. For instance, in run *Nounfull NF* with the second facet enabled, only 69 blogs out of 331 relevant blogs were retrieved and consequently, the MAP value for this run was low, as shown in Table 5.12.

	Results
numQueries	39
numRetrieved	3801
numRelevant	331
numRelevantRetrieved	69
map	0.0231

Table 5.12: Results for first submitted Run NF: Second facet - all Topics

In cases where the retrieval has been successful - which means that enough relevant blogs have been obtained for a topic, also good MAP results were achieved. An example of this is outlined in Table 5.13 which holds the results for the first facet for topic 1101.

re-indexed. A re-indexing of the whole dataset has not been possible due to the large number of blogs

Another example where the classification was successful was topic 1103, as shown in Table 5.14.

TOPIC	RUN	MAP	R-Precision	P@10
1101	NF	0.2590	0.4000	0.5000
1101	PF	0.2989	0.4571	0.6000
1101	SF	0.2221	0.4143	0.3000

Table 5.13: Results for third submitted Run SF: first Facet for Topic 1101

TOPIC	RUN	MAP	R-Precision	P@10
1103	NF	0.0525	0.1429	0.1000
1103	PF	0.2857	0.2857	0.2000
1103	SF	0.2381	0.2857	0.2000

Table 5.14: Results for third submitted Run SF: first Facet for Topic 1103

Compared to the summary statistics for the first facet "opinionated" , which is given in Table 5.15, one can see, that for this topic 1103, the achieved results were above average for the runs *punctfull* and *sentencefull*.

MAP1.	MAP2	MAP3	R-Prec1.	R-Prec2
0.4286	0.1667	0.0000	0.4286	0.2857

Table 5.15: Results for third submitted Run SF: first Facet for Topic 1103

In some cases, the proposed relevance ranking has revealed its drawbacks, reflected in low values for R-precision. In other experiments on the annotated subset of the Blogs08 dataset for instance, a classification accuracy of 75% has been achieved for style-based features only (see Section 5.4.3). Besides, on this annotated subset, with different features e.g. stems, an accuracy of 91% on blog post level has been achieved (see Section 5.5.2).

Comparison with Best Performing Systems

In the TREC Blog Distillation Task 2009, 9 groups participated whereas overall 29 runs have been submitted.

As suggested within this thesis, also most of the other groups used a two-stage approach to firstly identify feeds that were relevant to the given topics and secondly to determine for which facet inclination a blog was relevant.

The approach proposed within this thesis differs from most of the other approaches in respect to which parts of the Blogs08 collection have been indexed. For this approach, the feed components of the data collection have been indexed - yet most of the other approaches indexed only the permalinks components.

As an overall remark, the challenge organizers found that from the results it can be derived that *"the faceted blog distillation task has been particularly challenging to the participating groups"* [Macdonald et al., 2010a].

The organizers also observed that the general retrieval performance of all participating systems had at best been only average. As stated by the challenge organizers in their overview paper, this underpins the inherent difficulty of the Faceted Blog Distillation task.

Table 5.16 shows the best and median MAP values for the facets that had to be assessed within the TREC challenge, as published by the TREC organizers in [Macdonald et al., 2010a]:

Table 5.16: Best and Median MAP Values for the in TREC 2009 assessed Facets

	Facet	MAP
Best	Baseline	0.3617
Median	Baseline	0.1285
Best	Opinionated	0.2338
Median	Opinionated	0.0727
Best	Factual	0.2945
Median	Factual	0.0685
Best	Official	0.3167
Median	Official	0.0560
Best	Personal	0.2995
Median	Personal	0.0937
Best	Indepth	0.3489
Median	Indepth	0.0549
Best	Shallow	0.1906
Median	Shallow	0.0250

The best performing approaches have been by the groups ICTNET, USI, FEUP, uogTr, BIT, and buptpris___2009 and they used a variety of techniques that are shortly described here. Note that the description of the approaches is based on the TREC-2009 Blog Track Overview Paper [Macdonald et al., 2010a].

BIT applied a combination of several language models, FEUP used BM25 to

derive a baseline ranking and then, they exploited temporal information in blog posts for ranking so that the temporal information of a blog post either amplified (new post) or reduced (old post) a post's relevance score. ICTNET also applied a BM25 based blog post ranking and then they globally ranked the whole blogs. The approach proposed by buptpris___2009 consisted of a topic relevance model in combination with query expansion based on terms from the blog description and some topic fields. The USI group used fuzzy aggregation techniques to combine scores at blog post level into an agglomerative score for the whole blog. They also experimented with smoothing relevance scores based on content wise similarity of the retrieved blogs. The uogTr group used a Machine Learning approach based on a Voting Model as well as a learning-to-rank method, specifically the AdaRank to learn ranking models for the fact inclinations [Macdonald et al., 2010a].

Lessons Learned

In this topic independent content facet experiment, the quality of blogs has been assessed whereas the quality has been defined by a variety of binary content aspects. The blog quality assessment has been performed in the context of the TREC Blog Distillation Task 2009.

The proposed system consists of a plain text index extracted from the XML feeds only. A number of 680k of 1.3 Million blogs has been successfully indexed. This index has then been used to retrieve candidate blogs for the given topics. From the top 2500 result blog entries, the top 100 blogs have been identified according to the accumulated relevance score of the particular blog entries. The resulting blogs have been classified into the pre-defined binary content facets using a Support Vector Machine trained on a manually labeled subset of the TREC Blogs08 dataset.

Three runs have been conducted, whereas one run has been based on lexical features and the other two on stylometric features.

The best run has been based on stylometric features, sentence statistics, respectively, for the facet *personal*. Using the topic-independent and simple stylometric features, the fifth best results of all groups have been achieved. The results for the facet *opinionated* have been the second best results, whereas as features, also sentence statistics were exploited. Apparently, *the proposed simple stylometric features also enable to address certain quality aspects of blogs.*

5.6 Summary

In this chapter, several experiments have been outlined in which content facets have been assessed in traditional media, Online news, and blogs following the methodology proposed in Section 5.2. First of all, a content correlation experiment revealed that high quality traditional media correlate with selected blogs. Based on this insight, blogs have been classified into commonly agreed upon newspaper categories whereas the training set consisted of labeled news articles. This experiment revealed that that classification schemes from traditional media can be mapped to blogs. **Therefore, this chapter provides an answer to Research Question 1.**

In this chapter, two types of features have been proposed; with the constraint that the features should only be derived from the content of the media documents. More specifically, (i) lexical features, and (ii) stylometric features have been proposed. In a cross-topic experiment, it has been evaluated which of the feature types is better suited to assess topic oriented or topic independent content facets. This experiment revealed that for topic independent content facets, stylometric features serve best whereas for topic oriented facets, lexical features should be used if topics stay the same over training and test set. **This chapter consequently provides also an answer to Research Question 3.**

For the credibility content facet, a novel feature has been introduced; namely the quantity structure of a blog compared to the quantity structure of a credible news source. This thesis showed that based on this feature, blogs can be filtered out that exhibit a total different publishing behavior. Such blogs are for instance spam blogs or advertising blogs.

A selection of the proposed lexical and stylometric features has also be applied in the context of an international challenge, namely the TREC 2009 Blog Distillation Task. In this challenge, several content facets had to be assessed that represent quality aspects of blogs. This experiment corroborated that stylometric features can be used to assess the quality related content facets and therefore, also the quality of blogs.

Chapter 6

Content Facet Assessment in Web Content

"You affect the world by what you browse" (Tim Berners-Lee)

In this chapter, the insights gained in the content facet experiments in the media domain (see Chapter 5) are applied to a more general problem of assessing content facets in arbitrary Web content. The experiments for this have been conducted in the context of another international challenge, the ECML/PKDD Discovery challenge [Lex et al., 2010d]. This gives insights related to Research Question 4, *"If content facets and features can be identified and extracted from media content, can they be generalized to Web content?"*, because similar topic oriented and topic independent content facets are assessed in arbitrary Web hosts with both Bag-of-Words and style based features.

6.1 Introduction

The ECML/PKDD Discovery Challenge 2010[1] aimed to develop automatic methods to estimate the overall rank, quality, and importance of Web content. The goal of the challenge has been to support organizations to prioritize the gathering, storing and organization of Web pages.

[1] `http://www.ecmlpkdd2010.org/articles-mostra-2041-eng-discovery_challenge_2010.htm`

This challenge contributes to this dissertation in respect to the following scenario: for media agencies like the Austria Press Agency APA, it is important to always provide a comprehensive amount of useful information sources. In many cases, the information sources are identified manually by domain experts. Due to the importance of the Web, media agencies nowadays also provide and archive Web content. The rationale behind that is that they aim to provide broad research opportunities for journalists and information professionals. A manual and comprehensive identification of potentially useful Web content is yet costly. The availability of automatically derived content quality measures would facilitate the automatic identification of useful information sources as well as the maintenance of these sources.

The next section outlines the ECML/PKDD Discovery Challenge as well as the conducted experiments in more detail.

6.2 The ECML/PKDD Discovery Challenge

"It is a very sad thing that nowadays there is so little useless information." (Oscar Wilde)

In this section, the ECML/PKDD Discovery Challenge 2010 is described in detail. From a Web archive point of view, the usefulness of content obtained from web crawls is sometimes questionable, especially in respect to information quality. If quality measures or rankings would be available in addition to the content itself, the archival would be improved as it can be automatically decided whether it is worth to archive a particular Web content or not. The ECML/PKDD Discovery Challenge 2010 has addressed this in form of three different tasks. These tasks are described in the next section.

6.2.1 Tasks

The ECML Discovery Challenge has consisted of three tasks: (i) a classification task to assess the Web genre and information quality facets like neutrality, bias, and trustiness, (ii) an English quality task whereas the quality of a Web site has been measured as an aggregate function of its genre and its neutrality, bias and trustiness, and (iii) a multilingual quality task where the quality of German and French Web sites has to be assessed.

Task 1

The goal of Task 1 has been to classify English Web hosts into a set of categories: Web Spam, News/ Editorial, Commercial, Educational/Research, Discussion, Personal/Leisure, and to assess the level of neutrality, bias, and trustiness on a scale from 1 to 3 whereas 3 denotes normal and 1 problematic content. The result of Task 1 has consisted of a ranked list whereas the test hosts have been ranked by classifier confidence.

Task 2

The aim of Task 2 has been to measure the quality of the English Web hosts whereas the quality has been determined as an aggregate function of the host's genre, its neutrality, bias, and trustiness. The facets neutrality, bias, and trustiness cover the intrinsic content quality, as described by Huang et al. in [Huang et al., 1999]. The overall quality score has derived by combining the results retrieved in Task 1 according to the following rule:

```
utilityScore = 0;
if (News-Edit OR Educational) {
value = 5;
} else if (Discussion) {
value = 4;
} else if (Commercial OR Personal-Leisure) {
value = 3;
}
if (neutrality == 3) value += 2;
if (bias == 1) value -= 2;
if (trustworthiness == 3) value +=2;
```

The rationale behind this definition of quality has been that the challenge organizers defined quality with regard to the needs of an Internet archive. Therefore, the categories News and Educational have the highest quality. Also, the rule implies that quality content should exhibit trust, no bias, and neutrality. Consequently, Web Spam hosts have by default the lowest quality. The result of Task 2 has also consisted of a list ranked by classifier confidence.

Task 3

Task 3 has aimed at assessing the quality of German and French Web hosts since in the .eu domain, a lot of content is available in other languages than English. The focus in this task has been thereby on two major European languages, German and French. The quality of the German and French hosts has also derived using the above rule and as a result, a list ranked by classifier confidence is obtained.

6.2.2 Dataset and Features

The dataset for the Discovery Challenge 2010 has been based on a crawl of the .eu domain provided by the European Archive Foundation[2]. The dataset contains a collection of annotated Web hosts labeled by the Hungarian Academy of Sciences (English), European Archive Foundation (French) and L3S Hannover (German) [Benczur et al., 2010]. Table 1 shows the number of English training samples for each class: the dataset is in most cases highly imbalanced towards the positive class. Note that while the genre categories are mutually exclusive, the quality categories are not.

Table 6.1: ECML Challenge: Class Distribution

Category	Positive Samples [%]	Negative Samples [%]
WebSpam	4	96
News/Editorial	4.7	95.3
Educational/Research	43	57
Personal/Leisure	23.7	76.3
Commercial	45.4	54.6
Discussion	5.3	94.7
Bias	1.7	98.3
Neutrality	96.6	3.4
Trustworthiness	98.1	1.9

Features

In the ECML/PKDD dataset, different types of features have been provided. Most features were assessed on a per host level, only the natural language processing

[2] `http://datamining.sztaki.hu/?q=en/DiscoveryChallenge/`

features were available on a large set of sample pages. The features are described in more detail in the next paragraphs. Note that as dataset backend, feature vectors have been created for each feature set which then have been stored in an Apache Lucene [3] index. This resulted in an index size of approximately 19 GB.

Link Features The provided link based features were derived from the Web graph and were available on a per host level. The feature set contains features like the in-degree, the out-degree, the PageRank, the edge reciprocity, the assortativity coefficient, and the TrustRank, summing up to 176 features.

Content based Features The content based features have also been available on a per host level. This feature set contains features like the number of words in the homepage or the average length of the title. They were proposed by Castillo et al. in [Castillo et al., 2007] to detect Web spam based on content. In this experiment, all given content based features have been exploited, summing up to 95 features.

Natural Language Processing Features The Natural Language Processing (NLP) features have been available per URL in contrast to the other feature sets. They were processed by the LivingKnowledge project[4]. Included in this feature set are the counts for sentence, token, character, the count of various Part-of-Speech (POS) tags, etc. Therefore, these features cover style based properties. Generally, stylometric features are well suited for assessing quality facets like neutrality since they are inherently topic independent [Lex et al., 2010a, Lex et al., 2010c]. Note that all provided NLP features have been used, except the most common bigrams - since they were often null, resulting in 180 NLP features.

Term Frequencies This feature set consists of the host level aggregate term vectors of the most frequent terms. Note that the top 50,000 terms are considered after eliminating stop words. The term frequency is computed over an entire host while the document frequency is on page level. The term frequency and the document frequency have then be exploited to weight the features with TF-IDF.

[3] Apache Lucene, `http://lucene.apache.org/`, last accessed May 2011
[4] http://livingknowledge-project.eu/

6.2.3 Approach

In this approach towards content facet assessment in Web content, an ensemble classifier strategy has been applied to exploit all types of features that were provided by the challenge organizers.

Each classification task has been addressed as a binary classification strategy. More specifically, the test hosts have been classified into the positive versus the negative class using the different classifiers. Then, the achieved classification results have been combined based on a majority voting whereas the test hosts have been assigned to the winner with the maximum classifier confidence.

For the multi language quality task (Task 3), only the link based and content based features derived from the English training hosts have been considered. The training set for both the German and French hosts contained only a few annotated hosts. Therefore, also the English link based features for the multilingual quality task have been exploited since they are inherently language independent. Also, the English content based features have been used for two reasons: Firstly, such features have originally been proposed by Castillo et al. [Castillo et al., 2007] to detect Web spam. Since spam is typically not identified by language, the assumption was that the content based features can also be exploited over different languages. Secondly, the content features basically correspond to the stylometric features that have been proposed within these dissertation research. Therefore, they should also be applicable over related, indoeuropean languages.

Classifiers

For the ensemble based approach, three different classification algorithms have been used. Firstly, the implementation of a J48 decision tree [Hall et al., 2009] given in Weka[5] has been exploited whereas the parameters have been set to $C = 0.25$ and $M = 2$.

To compensate the imbalance in the category representation in the given training set, a filter based on Synthetic Minority Oversampling Technique (SMOTE) has been applied. In the SMOTE technique, artificial training samples are generated for the minority class based on the k nearest neighbors of a training item [Chawla et al., 2002]. Therefore, the minority class is oversampled exploiting the artificial training sam-

[5] `http://www.cs.waikato.ac.nz/ml/weka/`

ples. It is also worth mentioning that random sampling has also been applied, however, with not much effect. Note that the SMOTE implementation given in Weka [Hall et al., 2009] has been used. Regarding the parameterization of SMOTE, the number of nearest neighbors has been set to 5, the percentage to 100, and the random seed to 1. Additionally, the feature values have been normalized using a normalization filter from Weka.

Secondly, the Class-Feature-Centroid Classifier (CFC) (see Section 2.4.3) has been used, among other because within this thesis, it has already successfully be used for genre classification in English blogs (see Section 5.4.3).

Thirdly, a Support Vector Machine (SVM) based on LibLinear [Fan et al., 2008] has been applied since SVMs are among the best text classification algorithms and especially the LibLinear is known to be fast and efficient.

In this approach, the three classifiers have been used with different feature sets: On the term frequencies, the CFC algorithm has been applied since its highly discriminant abilities serves best in this setting. The CFC algorithms needs real terms to compute its discriminative weighting scheme and fortunately, the challenge organizers also provided a dictionary of the 50000 top terms. Therefore, this highly efficient algorithm could be used. The assumption has been that especially for topic driven categories like *News/Editorial* and *Educational/Research*, the CFC serves well.

On the link based and content based features, the J48 classifier has been applied with the SMOTE filter since cross-validation experiments on the training set revealed that this classifier deals best with the imbalance problem. Note that the J48 classifier has already been successfully applied to a similar problem of spam classification, as described by Castillo et al. in [Castillo et al., 2007] with the only difference that they used it as a base classifier for a cost-sensitive classifier.

In the ECML/PKDD experiments, also a cost sensitive classifier with J48 and similar parameters as described in [Castillo et al., 2007] has been evaluated, however the SMOTE based approach outperformed the cost sensitive classifier.

On the natural language processing features, the LibLinear implementation of a SVM has been used. This decision has been based on practical reasons only since in this case, there is a large amount of feature vectors (approx. 23M) because the natural language processing features were assessed on a page level - in contrast to all other features which were assessed on a per host level. Clearly, the LibLinear has

not been the best algorithm in this setting but it is very fast and highly efficient. Note that in order to determine the best performing cost parameter C, a grid search has been conducted that identified $C = 0.04$ as best parameter.

6.2.4 Results

The results for Task 1 are given in Table 6.2. Note that the evaluation has been conducted in terms of the evaluation metric Normalized Discounted Cumulated Gain (NDCG). The results for Task 1 revealed that the category *Educational/Research* achieves the best results in terms of NDCG. A manual examination of a number of test hosts identified that the categories *News/Editorial* and *Educational/Research* are quite hard to separate with the given features. A reason for this might be that both categories exhibit a similar writing style (factual, neutral, rather long and complex words) which results in similar content based and natural language processing features. Also, over both categories, similar terms are used.

Table 6.2: ECML Challenge: Results for Task 1

Category	NDCG
WebSpam	0.473
News/Editorial	0.416
Commercial	0.694
Educational/Research	0.688
Discussion	0.531
Personal/Leisure	0.583
Trustiness	0.397
Bias	0.540
Neutrality	0.51
Average	0.537

The results derived from Task 1 have then directly been used to compute the quality of the English test hosts and further to rank the English hosts by quality. The results for Task 2 are shown in Table 6.3: As one can see, the quality of the

Table 6.3: ECML Challenge: Results for Task 2

Language	NDCG
English	0.844

English hosts could be assessed quite well. This is clearly due to the fact that in Task 1, the category *Educational/Research* has been assessed with a rather good confidence. Note that this category has had a high influence in the final quality function. In the multilingual setting for Task 3, good results for the French and

Table 6.4: ECML Challenge: Results for Task 3

Language	NDCG
German	0.792
French	0.823

German hosts have been achieved, even though only the link based and content based features derived from the English training hosts have been used in this case. Note that the results for Task 3 are shown in Table 6.4. This reveals that the link and the content features are rather language independent, at least for indoeuropean languages, and robust.

6.2.5 Comparison with Best Performing Systems

In the ECML/PKDD Discovery Challenge 2010, five different systems have been submitted. The best performing approaches have been CAS, MADSPAM, and WXN. The overall winning approach has been submitted by the CAS group. Their approach has been based on the extraction of multi-scale features and a feature fusion strategy. From all available features, a joint features vector has been compiled that has been used for classification. As classifier, a bagging algorithm has been used with a C.45 decision tree as weak learner [Geng et al., 2010].

The MADSPAM approach won the English quality task. Their approach has been based on firstly training a ranking model on instance based features, and secondly, they used the available Web graph structure information to smooth the predictions that have been derived with the ranking model in the first step. As ranking algorithm, they exploited the RankBoost algorithm [Sokolov et al., 2010].

The WXN group employed a Wilcoxon based feature selection strategy in combination with ensemble classification as well as multiple binary classifiers. Note that this approach is actually quite similar to the approach that has been carried out within this dissertation research [Nikulin, 2010].

Table 6.5 shows the detailed results of the best performing systems.

Table 6.5: Results for Other Participants of the ECML/PKDD Discovery Challenge 2011

		Task1	EN	DE	FR	All
CAS	final	0.711657	0.935589	0.854482	0.833070	0.833700
MADSPAM	final	0.700951	0.923302	0.815732	0.836181	0.819041
	best	0.700951	0.923302	0.820690	0.845521	0.822616
WXN	final	0.660743	0.805396	0.775717	0.821081	0.765734
	best	0.704874	0.897536	0.801669	0.824412	0.807123

6.3 Summary

In the proposed approach towards the ECML/PKDD Discovery Challenge 2010, all provided features have been exploited in an ensemble classifier setting. Three different classifiers have been applied whereas each classifier has been trained on a different feature set. As a result, the proposed approach has achieved a runner-up position in the challenge.

The experiments also revealed that even if the NDCG is low for some categories like Web Spam, News/Editorial, and Bias, the quality of the Web hosts could be assessed with a high NDCG of 0.844 in the monolingual setting (English hosts), and a NDCG of 0.793 (German) and 0.823 (French) in the multilingual setting.

This experiment provides an answer to Research Question 4 since also stylometric features as well as lexical features have been proven applicable to assess quality related content facets in Web content.

Chapter 7

Conclusions and Outlook

"Whoever undertakes to set himself up as a judge of Truth and Knowledge is shipwrecked by the laughter of the gods." (Albert Einstein)

This dissertation has introduced methods to assess content facets both in traditional media as well as social media, blogs, respectively. The proposed content facets have been introduced as means to address the personal information needs of media consumers.

This chapter concludes this dissertation research and it provides a reflection on the goals set at the start of this research. The reflection is hereby carried out in form of an assessment in respect to the proposed research questions. This chapter also gives an outlook on possible links and directions for future work in the area of content facet assessment, and media analysis.

7.1 Self Assessment

In this section, the achievements are compared with the original research goals of this dissertation research. The research goals for this thesis have been listed in form of research questions in Section 1.1. In the following, the addressed research questions are reflected from a retrospective view. Special attention has been paid to reveal whether this dissertation research has answered the proposed research questions. Note that each research question is tackled in the order it has been originally posed.

Self Assessment for Research Question 1: *"How effectively can classification schemes from traditional media be mapped onto blogs? "*

Research Question 1 question has been answered in Section 5.4.2 in form of a cross-domain genre classification approach. The goal has been to map classification schemes from traditional media to news-related blogs.

In this context, a thematic classification has been performed on news related blogs using a model trained on a high quality news corpus. The labeled data from the news corpus has been exploited to classify blogs into commonly agreed upon newspaper categories. It is noteworthy that newspaper editors assigned the news articles to the used categories. Several text classifiers have been applied and the experiments revealed that **the classification schemes from traditional media can indeed be mapped onto blogs.**

This finding emphasized that content facets can be transferred from a high quality news corpus to news related blogs of informal and dynamic nature: consequently, both domains exhibit similar content-based characteristics. In other words, **this analysis revealed that quality news correlate topic wise with selected blogs.** To assess the content facet credibility, this correlation has been exploited to rank blogs by credibility based on the publishing behavior and the content similarity in comparison to credible news.

To analyze and visualize the common characteristics of news and blogs in more detail, the APA Labs Blog Trend Visualization module has been used (see Chapter 3). The Blog Trend Visualization returns, given a German search query, both news articles as well as blogs that are related to the search query and shows them over time. The benefit of the Blog Trend Visualization is that it visually provides an overview of how topics in news and blogs evolve over time. **From the Blog Trend Visualization it has been derived that topics in the news domain are reflected in the blogosphere, at least in one (German) language.**

To analyze not only the German part of the blogosphere, the APA Labs framework has been extended to support cross-language queries, namely queries in English, French, Spanish, and Italian. A cross-language content correlation evaluation revealed that **German news topics derived from the APA news repository are also reflected in English, French, Spanish and Italian blogs.**

Self Assessment for Research Question 2: *"Into which types of content facets can be blogs be categorized?"*

Research question 2 has been addressed and answered in Section 4.2 in form of a literature review. From this literature review, content facets have been divided into being **either topic oriented or topic independent content facets**.

The impact of this finding is that different types of content facets yield different types of features and algorithms. This issue has further been investigated in the context of Research Question 3.

Self Assessment for Research Question 3: *Research Question 3: "Which types of features are most suitable for detecting topic oriented facets and topic independent facets?"*

Research question 3 has been addressed by comparing the performance of standard bag-of-words features, more specifically lexical features, with a set of stylometric features that capture stylistic properties of documents.

In this dissertation, it has been shown that topic independent stylometric features are more suitable for the assessment of topic independent content facets. On the other hand, lexical features are better suited to assess topic oriented content facets as for example topic and genre. More specifically, for topic oriented facets or when topics stay the same in training and test set, lexical features shall be used. **This finding provides an answer to Research Question 3.**

Self Assessment for Research Question 4: *"If content facets and features can be identified and extracted from media content, can they be generalized to Web content?"*

This research question has been tackled in the context of an international challenge, namely the ECML/PKDD Discovery Challenge 2010. From the results achieved within the challenge, it can be derived that **both content facets as well as features from media content can be generalized to Web content.** For instance, in both media content as well as in Web content, the assessment of the content facet *genre* is definitely feasible.

Additionally, the quality of information is an important matter in social media and on the Web in general. Therefore, content facets that are related to information quality are feasible in both domains.

The experiments in the challenge also revealed that **stylometric features can be used to assess content facets in Web content as well** - and in combination with topic and language independent link features, they can even be used to assess the Web genre and the quality across different indoeuropean languages.

7.2 Open Questions

This section briefly gives an outlook on open questions of this dissertation research.

Within this dissertation research, the following open questions have been identified:

- How to solve genre classification?
- Are there more sophisticated features?
- User evaluation to identify whether facets help
- Performance difference between topic oriented and topic independent features
- The availability of a sufficiently large amount of training data for the content facets

The single-domain genre classification experiment outlined in Section 5.4.3 revealed that genre classification is still a hard problem, especially when the classification should be as topic independent as possible.

In Section 6.2, genre classification has also been performed in the context of Web classification. In this experiment, not only stylometric features have served as topic-independent features but also link features. For instance, for assessing the Web genre *WebSpam* the use of solely content features is definitely not sufficient - also literature reports [Abernethy et al., 2008, Castillo et al., 2007] that a combination of content features with link features yields much better results. This directly leads to the next open question: for some content facets as for example genre, the assumption is that more sophisticated features are needed. The ECML/PKDD Discovery Challenge 2010 revealed that for instance, the genres News/Educational and Research are hard to separate solely based on the used terms (lexical features) or on stylometric features. The use of more sophisticated features as for example semantic features might result in a performance increase.

The next open question is the feasibility of the proposed content facets in media from the user's point of view. For this, a detailed user evaluation would be needed in order to identify whether the content facets help at all or, which content facets are feasible and which not. To a certain extend, some conclusions can be drawn for the content facets that have been implemented in the APA Labs framework. The Austrian Press Agency APA provides on the Web site of APA Labs [1] the possibility to rate the module of APA Labs on a five point Likert scale and to give feedback to them. Since the proposed modules and consequently, the actually implemented content facets, have been well received, it seems reasonable to assume that these content facets are helpful. However, the feasibility aspect of the proposed content facets needs to be investigated in more detail.

Another open question is the performance difference between topic oriented and topic independent features. Apparently, with topic oriented features, the maximum achievable performance is higher than in the topic independent case - independent of classification task or whether enough training data are available.

This directly leads to the next open point, the availability of training data. Annotating training data is a labor intensive task and it requires domain knowledge as well as temporal resources. In this thesis, many of the training data sets have been manually annotated. Naturally, a more efficient creation of training data would be much more feasible; since this has not been focus of this dissertation research, this directly leads to directions for future work in this respect.

7.3 Impact

With the fact that more and more content is available in the media domain, the challenges addressed in this dissertation are highly relevant; people need to become media analysts themselves to cope with the information overload.

The impact of this dissertation research is that the suggested content facets provide a solution to support media consumers to filter traditional and social media content with respect to their personal information need. These content facets can be applied in a manifold of applications as for example faceted search systems or information filtering applications. The proposed APA Labs framework represents such a faceted search system.

[1] APA Labs, `www.apa.at/labs`, last accessed March 2011

Besides, this dissertation suggests methods to assess quality aspects of traditional media and social media directly from content. As a consequence, traditional and social media become comparable in terms of quality. This is highly beneficial for both media consumers as well as media analysts.

Naturally, the proposed content facets can be useful in other domains as well; especially the quality related content facets assessed with stylometric features can be beneficial to judge the value of any content in general.

As the experiment in the context of the ECML/PKDD Discovery Challenge 2010 showed, for certain applications it is feasible to combine several content facets into a single quality score. Providing such a score in addition to the content itself can support users to identify whether they should spend time to examine the content or not. Besides, based on such a score, low quality content can be automatically filtered out.

7.4 Future Work

This section highlights directions for future work in the field of content facet assessment, feature engineering for content facet assessment, and the integration of content facets in faceted search systems.

Typically, the performance of a classifier is closely related to the number and quality of training data. Therefore, an interesting research opportunity would be to exploit more efficient means to create training data for the content facets. For instance, Active Learning or Co-training could be used to efficiently compile new training data.

Another interesting research opportunity would be to represent content wise *relations* as content facets. More specifically, the information who, what, where, and when something has happened can be beneficial in addition to the content itself. Within this dissertation research, to a certain extent, such relational facets have been assessed in traditional media - yet on a per document basis and in form of a Named Entity extraction. A more complex analysis of such relational facets on for instance sentence or paragraph level as well as the incorporation of semantic aspects would enable the user to search in media content by relations.

As mentioned in Section 7.2, the applicability of the proposed content facets in faceted search system, as well as the perceived usefulness of the content facets is

an interesting direction for future work. This can be addressed in form of extensive user studies and evaluations.

As shown in Section 6.2, content facets can be used to assess the quality and credibility of information. Using content facets e.g. to determine the quality of learning resources in the Technology Enhanced Learning (TEL) community or to provide access to learning resources based on a user's personal information need would also be an interesting research direction.

Also, the design of a feasible user interface especially for the content facets in blogs and social media would be worth investigating in the future. Such a user interface could e.g. enable to share and recommend resources to other users or to give feedback to the content facet assessment. This could even result in providing users with the opportunity to create classification models that meet their personal model of the particular content facet.

Bibliography

[Abel et al., 2010] Abel, F., Henze, N., and Krause, D. (2010). Optimizing search and ranking in folksonomy systems by exploiting context information. In Aalst, W., Mylopoulos, J., Rosemann, M., Shaw, M. J., Szyperski, C., Cordeiro, J., and Filipe, J., editors, *Web Information Systems and Technologies*, volume 45 of *Lecture Notes in Business Information Processing*, pages 113–127. Springer Berlin Heidelberg.

[Abernethy et al., 2008] Abernethy, J., Chapelle, O., and Castillo, C. (2008). Web spam identification through content and hyperlinks. In *Proceedings of the 4th international workshop on Adversarial information retrieval on the web*, AIRWeb '08, pages 41–44, New York, NY, USA. ACM.

[Abramowicz, 2003] Abramowicz, W. (2003). *Knowledge-Based Information Retrieval and Filtering from the Web (The Kluwer International Series in Engineering and Computer Science, Secs 746)*. Kluwer Academic Publishers, Norwell, MA, USA.

[Agarwal and Liu, 2008] Agarwal, N. and Liu, H. (2008). Blogosphere: Research issues, tools, and applications. *SIGKDD Explorations*.

[Agichtein et al., 2008] Agichtein, E., Castillo, C., Donato, D., Gionis, A., and Mishne, G. (2008). Finding high-quality content in social media. In *Proceedings of the international Conference on Web Search and Web Data Mining (WSDM)*.

[Aha et al., 1991] Aha, D., Kibler, D., and Albert, M. (1991). Instance-based learning algorithms. *Journal of Machine Learning Research*, pages 139–149.

[Akamine et al., 2009] Akamine, S., Kawahara, D., Kato, Y., Nakagawa, T., Inui, K., Kurohashi, S., and Kidawara, Y. (2009). Wisdom: a web information credi-

bility analysis system. In *Proceedings of the ACL-IJCNLP 2009 Software Demonstrations*, ACLDemos '09, pages 1–4, Stroudsburg, PA, USA. Association for Computational Linguistics.

[An et al., 2004] An, A., Huang, Y., Huang, X., and Cercone, N. (2004). Feature selection with rough sets for web page classification, 2002. supported by natural. In *Sciences and Engineering Research Council (NSERC) of Ontario, Canada and the Institute for Robotics and Intelligent Systems (IRIS*.

[Anderson, 2006] Anderson, C. (2006). *The Long Tail: How endless choice is creating unlimited demand*. Random House Business Books.

[Antia, 2007] Antia, B. E. (2007). *Indeterminacy in Terminology and LSP*. John Benjamins.

[APA-IT, 2011] APA-IT (2011). Apa-it informations technologie gmbh powersearch database homepage. `http://www.apa-it.at/cms/it/DE/loesungen.html?channel=CH0265&doc=CMS1147877187557`. [Online; accessed 03-January-2011].

[Araujo and Martinez-Romo, 2010] Araujo, L. and Martinez-Romo, J. (2010). Web spam detection: new classification features based on qualified link analysis and language models. *Trans. Info. For. Sec.*, 5:581–590.

[Arlot and Celisse, 2004] Arlot, S. and Celisse, A. (2004). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 15:551–564.

[Atzmüller and Landl, 2009] Atzmüller, P. and Landl, G. (2009). Semantic enrichment and added metadata - Examples of efficient usage in an industrial environment. *World Patent Information*, pages 89–96.

[Baeza-Yates and Ribeiro-Neto, 1999] Baeza-Yates, R. A. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

[Belkin and Croft, 1992] Belkin, N. J. and Croft, W. B. (1992). Information filtering and information retrieval: two sides of the same coin? *Commun. ACM*, 35:29–38.

[Ben-Yitzhak et al., 2008] Ben-Yitzhak, O., Golbandi, N., Har'El, N., Lempel, R., Neumann, A., Ofek-Koifman, S., Sheinwald, D., Shekita, E., Sznajder, B., and Yogev, S. (2008). Beyond basic faceted search. In *Proceedings of the international conference on Web search and web data mining*, WSDM '08, pages 33–44, New York, NY, USA. ACM.

[Benczur et al., 2010] Benczur, A., Castillo, C., Erdelyi, M., Gyöngyi, Z., Masanes, J., and Matthews, M. (2010). ECML/PKDD 2010 Discovery Challenge Data Set. Crawled by the European Archive Foundation.

[Bishop, 2008] Bishop, C. M. (2008). *Pattern Recognition and Machine Learning*. Number 978-0387310732. Springer-Verlag New York Inc.

[Blum and Langley, 1997] Blum, A. L. and Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artif. Intell.*, 97:245–271.

[Blumenstock, 2008] Blumenstock, J. E. (2008). Size matters: word count as a measure of quality on wikipedia. In *Proceeding of the 17th international conference on World Wide Web*.

[Börkur and van Zwol Roelof, 2010] Börkur, S. and van Zwol Roelof (2010). Tagexplorer: Faceted browsing of flickr photos. Technical report, Yahoo! Labs Technical Report.

[Boström, 2006] Boström, F. (2006). Personal information retrieval.

[Broder and Maarek, 2006] Broder, A. Z. and Maarek, Y. S. (2006). Sigir'2006 workshop on faceted search, call for participation. http://sites.google.com/site/facetedsearch/.

[Buckley and Voorhees, 2000] Buckley, C. and Voorhees, E. M. (2000). Evaluating evaluation measure stability. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 33–40, New York, NY, USA. ACM.

[Burges, 1998] Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:121–167.

[Bush, 1945] Bush, V. (1945). As we may think. *The Atlantic Monthly*, 176(1):101–108.

[Carmel et al., 2009] Carmel, D., Roitman, H., and Zwerdling, N. (2009). Enhancing cluster labeling using wikipedia. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 139–146, New York, NY, USA. ACM.

[Carmel et al., 2006] Carmel, D., Yom-Tov, E., Darlow, A., and Pelleg, D. (2006). What makes a query difficult? In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 390–397, New York, NY, USA. ACM.

[Castillo et al., 2007] Castillo, C., Donato, Murdock, V., and Silvestri, F. (2007). Know your neighbors: Web Spam Detection using the Web Topology. In *Proceedings of SIGIR*. ACM.

[Cavnar and Trenkle, 1994] Cavnar, W. B. and Trenkle, J. M. (1994). N-gram-based text categorization. In *Proceedings of SDAIR-94 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175.

[Chan et al., 2007] Chan, J., Poon, J., and Koprinska, I. (2007). Enhancing the performance of semi-supervised classification algorithms with bridging. In *FLAIRS Conference'07*, pages 580–585.

[Chang and Lin, 2001] Chang, C.-C. and Lin, C.-J. (2001). *LIBSVM: a library for support vector machines*. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[Chawla et al., 2001] Chawla, N., Eschrich, S., and Hall, L. (2001). Creating ensembles of classifiers. In *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on Data Mining*, pages 580–581.

[Chawla et al., 2002] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16:321–357.

[Chen and Choi, 2008] Chen, G. and Choi, B. (2008). Web page genre classification. In *Proceedings of the 2008 ACM symposium on Applied computing*, SAC '08, pages 2353–2357, New York, NY, USA. ACM.

[Chesley et al., 2005] Chesley, P., Vincent, B., Xu, L., and Srihari, R. (2005). Using verbs and adjectives to automatically classify blog sentiment. In *AAAi*.

[Cleverdon, 1970] Cleverdon, C. W. (1970). The effect of variations in relevance assessments in comparative experimental tests of index languages. Technical report, Cranfield Institute of Technology.

[Cleverdon et al., 1966] Cleverdon, C. W., Mills, J., and Keen, M. (1966). Factors determining the performance of indexing systems. *Design*.

[Cortes and Vapnik, 1995] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20:273–297. 10.1023/A:1022627411411.

[Coyotl-Morales et al., 2006] Coyotl-Morales, R. M., Villasenor-Pineda, L., Montes-Y-Gomez, M., and Rosso, P. (2006). Authorship attribution using word sequences. *Proc of the 11th Iberoamerican Congress on Pattern Recognition CIARP 2006*, LNCS 422:844–853.

[Crestani, 1998] Crestani, F. (1998). Sonification of an information retrieval environment: Design issues. In *Proceedings of the First International Forum on Multimedia and Image Processing*.

[Cutting et al., 1992] Cutting, D. R., Karger, D. R., Pedersen, J. O., and Tukey, J. W. (1992). Scatter/gather: a cluster-based approach to browsing large document collections. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '92, pages 318–329, New York, NY, USA. ACM.

[Dakka and Ipeirotis, 2008] Dakka, W. and Ipeirotis, P. G. (2008). Automatic extraction of useful facet hierarchies from text databases. In *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering*, pages 466–475, Washington, DC, USA. IEEE Computer Society.

[de Winter and de Rijke, 2007] de Winter, W. and de Rijke, M. (2007). Identifying facets in query-biased sets of blog posts. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, pages 251–254.

[Decrem, 2006] Decrem, B. (2006). Introducing flock beta 1.

[Delbru, 2010] Delbru, R. (2010). *Searching Web Data: an Entity Retrieval Model*. PhD thesis, Digital Enterprise Research Institute.

[Dennis et al., 2002] Dennis, S., Bruza, P., and Mcarthur, R. (2002). Web searching: A process-oriented experimental study of three interactive search paradigms. *JASIST*, (2):120–133.

[Drezner and Farrell, 2004] Drezner, D. and Farrell, H. (2004). The power and politics of blogs. In *Proceedings of the American Political Science Association Conference (APSA)*.

[Duan and Keerthi, 2005] Duan, K.-B. and Keerthi, S. S. (2005). Which is the best multiclass svm method? an empirical study. In Oza, N. C., Polikar, R., Kittler, J., and Roli, F., editors, *Multiple Classifier Systems*, volume 3541 of *Lecture Notes in Computer Science*, pages 732–760. Springer Berlin Heidelberg.

[Duda et al., 2001] Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2 edition.

[Dumais et al., 2003] Dumais, S., Cutrell, E., Cadiz, J., Jancke, G., Sarin, R., and Robbins, D. C. (2003). Stuff i've seen: a system for personal information retrieval and re-use. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, pages 72–79, New York, NY, USA. ACM.

[Dunning, 1994] Dunning, T. (1994). Statistical identification of languages. Technical Report MCCS.

[Edmunds and Morris, 2000] Edmunds, A. and Morris, A. (2000). The problem of information overload in business organisations: a review of the literature. *International Journal of Information Management*, 20(1):17–28.

[English et al., 2002a] English, J., Hearst, M., Sinha, R., Swearingen, K., and Yee, K.-P. (2002a). Flexible search and navigation using faceted metadata. Technical report, University of Berkeley.

[English et al., 2002b] English, J., Hearst, M., Sinha, R., Swearingen, K., and Yee, K.-P. (2002b). Hierarchical faceted metadata in site search interfaces. In *CHI '02 extended abstracts on Human factors in computing systems*, CHI '02, pages 628–639, New York, NY, USA. ACM.

[Eppler, 2003] Eppler, M. (2003). *Managing information quality: Increasing the value of information in knowledge-intensive products and processes ; with 26 tables*. Springer, Berlin.

[Errecalde et al., 2008] Errecalde, M. L., Ingaramo, D., and Rosso, P. (2008). Proximity estimation and hardness of short-text corpora. In *Proceedings of the 2008 19th International Conference on Database and Expert Systems Application*, pages 15–19, Washington, DC, USA. IEEE Computer Society.

[Fan and Friedman, 2008] Fan, J. and Friedman, C. (2008). Word sense disambiguation via semantic type classification. In *Proceedings of the American Medical Informatics Association (AMIA) Symposium*, pages 177–181.

[Fan et al., 2008] Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, pages 1871–1874.

[Farmer, 2010] Farmer, D. (2010). Study projects nearly 45-fold annual data growth by 2020. `http://www.emc.com/about/news/press/2010/20100504-01.htm`. [Online; accessed 28-December-2010].

[Farooq et al., 2007] Farooq, U., Kannampallil, T. G., Song, Y., Ganoe, C. H., Carroll, J. M., and Giles, L. (2007). Evaluating tagging behavior in social bookmarking systems: metrics and design heuristics. In *Proceedings of the 2007 international ACM conference on Supporting group work*, GROUP '07, pages 351–360, New York, NY, USA. ACM.

[Feldman, 2004] Feldman, S. (2004). The high cost of not finding information. electronic magazine.

[Fielding and Taylor, 2002] Fielding, R. T. and Taylor, R. N. (2002). Principled design of the modern web architecture. *ACM Transactions of Internet Technology*, 2:115–150.

[Finn and Kushmerick, 2006] Finn, A. and Kushmerick, N. (2006). Learning to classify documents according to genre: Special topic section on computational analysis of style. *J. Am. Soc. Inf. Sci. Technol.*, 57:1506–1518.

[Fogg et al., 2001] Fogg, B., Marshall, J., Kameda, T., Solomon, J., Rangnekar, A., Boyd, J., and Brown, B. (2001). Web credibility research: a method for online experiments and early study results. In *CHI '01 extended abstracts on Human factors in computing systems*, CHI EA '01, pages 295–296, New York, NY, USA. ACM.

[Freund, 1995] Freund, Y. (1995). Boosting a weak learning algorithm by majority. *Inf. Comput.*, 121:256–285.

[Freund and Schapire, 1996] Freund, Y. and Schapire, R. E. (1996). Experiments with a New Boosting Algorithm. In *In Proceedings of the Thirteenth International Conference on Machine Learning*, pages 148–156.

[Fujimura et al., 2006] Fujimura, K., Toda, H., Inoue, T., Hiroshima, N., Kataoka, R., and Sugizaki, M. (2006). Blogranger - a multi-faceted blog search engine. In *In Proc. 3rd Annual WWE.*

[Furnas et al., 1987] Furnas, G. W., Landauer, T. K., Gomez, L. M., and Dumais, S. T. (1987). The vocabulary problem in human-system communication. *Communications of the ACM*, 30:964–971.

[Garfield et al., 1984] Garfield, E., Sher, I., and Torpie, R. (1984). The use of citation data in writing the history of science. *Journal of the American Society for Information Science and Technology.*

[Garner, 2007] Garner, J. (2007). Star searcher. Available Online.

[Geng et al., 2010] Geng, G.-G., Jin, X.-B., Zhang, X.-C., and Zhang, D. (2010). Evaluating web content quality via multi-scale features. In *Proceedings of ECML/PKDD Discovery Challenge, Available online.*

[GiHong and SangKi, 2008] GiHong, K. and SangKi, H. (2008). A study of online (digital) reputation in blogosphere based on relationship and activity. In *Proceedings of the International Conference on Cyberworlds*.

[Granitzer, 2004] Granitzer, M. (2004). Hierarchical text classification using methods from machine learning. Master's thesis, Graz University of Technology.

[Granitzer, 2006] Granitzer, M. (2006). *KnowMiner: Konzeption und Entwicklung eines generischen Wissenserschliessungsframeworks*. PhD thesis, University of Technology Graz.

[Granitzer et al., 2010] Granitzer, M., Kienreich, W., Sabol, V., and Lex, E. (2010). Knowledge Relationship Discovery and Visually Enhanced Access for the Media Domain. In *Medien - Wissen - Bildung: Explosionen visualisierter und Kollaborativer Wissensrume*, pages 46–58. Innsbruck University Press.

[Greene and G.M, 1971] Greene, B. and G.M, R. (1971). Automatic grammatical tagging of english. Technical report, Department of Linguistics, Brown University.

[Grossman, 2006] Grossman, L. (2006). Time's person of the year: You.

[Gruhl et al., 2004] Gruhl, D., Liben-Nowell, D., Guha, R., and Tomkins, A. (2004). Propagation of trust and distrust. In *Proceedings of the 13th international conference on World Wide Web*.

[Guan et al., 2009] Guan, H., Zhou, J., and Guo, M. (2009). A class-feature-centroid classifier for text categorization. In *Proceedings of the International Conference on World Wide Web (WWW)*.

[Guha and Tomkins, 2004] Guha, R., K. R. R. P. and Tomkins, A. (2004). Propagation of trust and distrust. In *Proceedings of the 13th international conference on World Wide Web (WWW'04)*, pages 403–412.

[Guyon and Elisseeff, 2003] Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3.

[Hall et al., 2009] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11 (1).

[Han and Karypis, 2000a] Han, E.-H. and Karypis, G. (2000a). Centroid-based document classification: Analysis and experimental results. In *Proceedings of the European Conferences on Principles of Data Mining and Knowledge Discovery (PKDD)*, pages 424–431.

[Han and Karypis, 2000b] Han, E.-H. and Karypis, G. (2000b). Centroid-based document classification: Analysis and experimental results. In *PKDD '00: Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery*.

[Harman, 1992] Harman, D. (1992). Overview of the first text retrieval conference (trec-1). In *TREC*, pages 1–20.

[Harris, 1997] Harris, R. (1997). Evaluating internet research sources. In *Home Page. 17 Nov. 1997. Vanguard University*.

[Hass et al., 2008] Hass, B. H., Kilian, T., and Walsh, G., editors (2008). *Web 2.0: Neue Perspektiven für Marketing und Medien*. Springer-11775 /Dig. Serial]. Springer-Verlag Berlin Heidelberg, Berlin, Heidelberg.

[Hearst, 2005] Hearst, M. (2005). Faceted Metadata in Search Interfaces (Invited Talk). Available online. `http://courses.ischool.berkeley.edu/i141/f05/lectures/hearst-facets.pdf`.

[Hearst, 2006a] Hearst, M. (2006a). Design recommendations for hierarchical faceted search interfaces. *ACM SIGIR Workshop on Faceted Search*.

[Hearst, 2009a] Hearst, M. (2009a). Faceted metadata for site navigation and search. Available online. `http://www.slideshare.net/marti_hearst/faceted-metadata-for-site-navigation-and-search`.

[Hearst, 2006b] Hearst, M. A. (2006b). Clustering versus faceted categories for information exploration. *Commun. ACM*, 49:59–61.

[Hearst, 2009b] Hearst, M. A. (2009b). *Search User Interfaces*. Cambridge University Press.

[Hearst et al., 2008] Hearst, M. A., Hurst, M., and Dumais, S. T. (2008). What should blog search look like? In *SSM '08: Proceeding of the 2008 ACM workshop on Search in social media*.

[Hearst et al., 1995] Hearst, M. A., Karger, D. R., and Pedersen, J. O. (1995). Scatter/gather as a tool for the navigation of retrieval results. In *Working Notes of the AAAI Fall Symposium on AI Applications in Knowledge Navigation and Retrieval, Cambridge, MA*.

[Hearst and Stoica, 2009] Hearst, M. A. and Stoica, E. (2009). Nlp support for faceted navigation in scholarly collections. In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, NLPIR4DL '09, pages 62–70, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Hoorn, 2010] Hoorn, Johan F., v. W. T. D. (2010). *Web Intelligence for the Assessment of Information Quality: Credibility, Correctness, and Readability, Web Intelligence and Intelligent Agents*. InTech.

[Hsin-Yih Lin et al., 2008] Hsin-Yih Lin, K., Yang, C., and Chen, H.-H. (2008). Emotion classification of online news articles from the reader's perspective. In *WI-IAT '08: Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pages 220–226, Washington, DC, USA. IEEE Computer Society.

[Hsu et al., 2003] Hsu, C., Chang, C., and Lin, C.-J. (2003). A Practical Guide to Support Vector Classification. *Bioinformatics*, 1:1–15.

[Huang et al., 1999] Huang, K.-T., Lee, Y. W., and Wang, R. Y. (1999). *Quality information and knowledge*. Prentice Hall PTR, Upper Saddle River, NJ, USA.

[Huang et al., 2006a] Huang, S., Sun, J.-T., Wang, X., Zeng, H.-J., and Chen, Z. (2006a). Subjectivity categorization of weblog with part-of-speech based smoothing. *International Conference on Data Mining*.

[Huang et al., 2006b] Huang, S., Sun, J.-T., Wang, X., Zeng, H.-J., and Chen, Z. (2006b). Subjectivity categorization of weblog with part-of-speech based smoothing. In *Proceedings of the Sixth International Conference on Data Mining*, ICDM '06, pages 285–294, Washington, DC, USA. IEEE Computer Society.

[Iding et al., 2008] Iding, M., Auernheimer, B., and Crosby, M. E. (2008). A metacognitive approach to credibility determination. In *Proc. of the 2nd Workshop on Information Credibility on the Web.*

[Ikeda et al., 2008] Ikeda, D., Takamura, H., and Okumura, M. (2008). Semi-supervised learning for blog classification. In *Proceedings of Twenty-Third AAAI Conf. on Artificial Intelligence.*

[Iwayama et al., 2005] Iwayama, M., Fujii, A., and Kando, N. (2005). Overview of classification subtask at ntcir-5 patent retrieval task. In *Proceedings of NTCIR-5 Workshop Meeting.*

[Järvelin and Kekäläinen, 2000] Järvelin, K. and Kekäläinen, J. (2000). IR Evaluation Methods for Retrieving Highly Relevant Documents. In *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 41–48, New York, NY, USA. ACM.

[Jatowt et al., 2009] Jatowt, A., Kanazawa, K., Oyama, S., and Tanaka, K. (2009). Supporting analysis of future-related information in news archives and the web. In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, JCDL '09, pages 115–124, New York, NY, USA. ACM.

[Jiang and Argamon, 2008a] Jiang, M. and Argamon, S. (2008a). Exploing subjectivity analysis in blogs to improve political leaning categorization. In *Proceedings of SIGIR.*

[Jiang and Argamon, 2008b] Jiang, M. and Argamon, S. (2008b). Finding political blogs and their political leanings. In *Text Mining 2008, Workshop at the SIAM International Conference on Data Mining.*

[Joachims, 1998] Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *European Conference on Machine Learning (ECML)*, pages 137–142, Berlin. Springer.

[Joachims, 2002] Joachims, T. (2002). *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms.* Kluwer Academic Publishers, Norwell, MA, USA.

[Jones et al., 2000] Jones, K. S., Walker, S., and Robertson, S. E. (2000). A probabilistic model of information retrieval: development and comparative experiments. *Inf. Process. Manage.*, 36(6).

[Juffinger et al., 2009a] Juffinger, A., Granitzer, M., and Lex, E. (2009a). Blog credibility ranking by exploiting verified content. In *Proceedings of the Workshop on Information Credibility on the Web (WICOW) in conjunction with WWW'2009*, pages 51–58, New York, NY, USA. ACM.

[Juffinger et al., 2009b] Juffinger, A., Kern, R., and Granitzer, M. (2009b). Cross-language retrieval based on wikipedia statistics. In *Proceedings of the 9th Cross-language evaluation forum conference on Evaluating systems for multilingual and multimodal information access*, CLEF'08, pages 155–162, Berlin, Heidelberg. Springer-Verlag.

[Juffinger and Lex, 2009] Juffinger, A. and Lex, E. (2009). Crosslanguage blog mining and trend visualisation. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 1149–1150, New York, NY, USA. ACM.

[Juffinger et al., 2009c] Juffinger, A., Neidhart, T., Granitzer, M., Kern, R., Weichselbraun, A., Wohlgenannt, G., and Scharl, A. (2009c). Distributed web 2.0 crawling for ontology evolution. *Journal of Digital Information Management,*, pages 114–119.

[Kale et al., 2007] Kale, A., Karandikar, A., Kolari, P., Java, A., Finin, T., and Joshi, A. (2007). Modeling trust and influence in the blogoshpere using link polarity. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*.

[Kang, 2010] Kang, M. (2010). Measuring social media credibility: A study on a measure of blog credibility. Institute for Public Relations, Winner of the 2009 Ketchum Award. Available online, `http://www.instituteforpr.org/topics/measuring-blog-credibility/`.

[Karlgren and Cutting, 1994] Karlgren, J. and Cutting, D. (1994). Recognizing text genres with simple metrics using discriminant analysis. In *Proceedings of the 15th conference on Computational linguistics*, pages 1071–1075.

[Katerattanakul and Siau, 1999] Katerattanakul, P. and Siau, K. (1999). Measuring information quality of web sites: development of an instrument. In *Proceedings of the 20th international conference on Information Systems*, ICIS '99, pages 279–285, Atlanta, GA, USA. Association for Information Systems.

[Kearns, 1988] Kearns, M. (1988). Thoughts on hypothesis boosting. Unpublished manuscript.

[Kienreich et al., 2008] Kienreich, W., Lex, E., and Seifert, C. (2008). Apa labs: an experimental web-based platform for the retrieval and analysis of news articles. In *Proceedings of ICADIWT*.

[Kim et al., 2006] Kim, S.-B., Han, K.-S., Rim, H.-C., and Myaeng, S. H. (2006). Some effective techniques for naive bayes text classification. *IEEE Trans. on Knowl. and Data Eng.*, 18:1457–1466.

[Kleinberg, 1999] Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46.

[Klemm et al., 2001a] Klemm, Iding, M., and Speitel, T. (2001a). Do scientists and teachers agree on the credibility of media information sources? *International Journal of Instructional Media*, 28.

[Klemm et al., 2001b] Klemm, E. B., Iding, M., and Speitel, T. (2001b). Do scientists and teachers agree on the credibility of media information sources? *International Journal of Instructional Media*, 28.

[Koerner, 2009] Koerner, C. (2009). The motivation behind tagging. In *ACM Student Research Competition, Hypertext 2009*.

[Koerner et al., 2010] Koerner, C., Benz, D., Hotho, A., Strohmaier, M., and Stum, G. (2010). Stop thinking, start tagging: tag semantics emerge from collaborative verbosity. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 521–530, New York, NY, USA. ACM.

[Kohavi and Provost, 1998] Kohavi, R. and Provost, F. (1998). Glossary of terms. *Journal of Machine Learning*, 30:271–274.

[Konchady, 2008] Konchady, M. (2008). *Building Search Applications: Lucene, Lingpipe, and Gate*. Musée d'art contemporain de Montréal.

[Koppel and Schler, 2004] Koppel, M. and Schler, J. (2004). Authorship verification as a one-class classification problem. In *Proceedings of the twenty-first international conference on Machine learning*, ICML '04, pages 62–, New York, NY, USA. ACM.

[Kotsiantis, 2007] Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. *Informatica*, 31(3):249–268.

[Kowalkiewicz et al., 2006] Kowalkiewicz, M., Orlowska, M. E., Kaczmarek, T., and Abramowicz, W. (2006). Robust web content extraction. In *Proceedings of the 15th International Conference on World Wide Web (WWW'06*.

[Kullback and Leibler, 1951] Kullback, S. and Leibler, R. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86.

[Kuo et al., 2007] Kuo, B. Y.-L., Hentrich, T., Good, B. M. ., and Wilkinson, M. D. (2007). Tag clouds for summarizing web search results. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 1203–1204, New York, NY, USA. ACM.

[Langley, 1995] Langley, P. (1995). *Elements of machine learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[Langville and Meyer, 2006] Langville, A. N. and Meyer, C. D. (2006). *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Princeton University Press.

[Lee and Myaeng, 2002] Lee, Y.-B. and Myaeng, S. H. (2002). Text genre classification with genre-revealing and subject-revealing features. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '02, pages 145–150, New York, NY, USA. ACM.

[Lenhart and Fox, 2006] Lenhart, A. and Fox, S. (2006). Bloggers: A portrait of the internet's news storytellers.

[Lex et al., 2009a] Lex, E., Granitzer, M., and Juffinger, A. (2009a). Facet classification of blogs: Know-center at the trec 2009 blog distillation task. In *In Proceedings of the 18th Text REtrieval Conference.*

[Lex et al., 2010a] Lex, E., Granitzer, M., Muhr, M., and Juffinger, A. (2010a). Stylometric features for emotion level classification in news related blogs. In *Proceedings of the 9th RIAO Conference (RIAO 2010).*

[Lex et al., 2010b] Lex, E., Juffinger, A., and Granitzer, M. (2010b). A comparison of stylometric and lexical features for web genre classification and emotion classification in blogs. *Database and Expert Systems Applications, International Workshop on*, 0:10–14.

[Lex et al., 2010c] Lex, E., Juffinger, A., and Granitzer, M. (2010c). Objectivity classification in online media. In *HT '10: Proceedings of the 21st ACM conference on Hypertext and hypermedia*, pages 293–294, New York, NY, USA. ACM.

[Lex et al., 2010d] Lex, E., Khan, I., Bischof, H., and Granitzer, M. (2010d). Assessing the quality of web content. In *Proceedings of ECML/PKDD Discovery Challenge, Available online.*

[Lex et al., 2009b] Lex, E., Seifert, C., Granitzer, M., and Juffinger, A. (2009b). Automated blog classification: a cross domain approach. In *Proceedings of IADIS International Conference WWWInternet*. IADIS.

[Lex et al., 2009c] Lex, E., Seifert, C., Granitzer, M., and Juffinger, A. (2009c). Cross-domain classification: Trade-off between complexity and accuracy. In *Proceedings of the 4th International Conference for Internet Technology and Secured Transactions (ICITST).*

[Lex et al., 2010e] Lex, E., Seifert, C., Juffinger, A., and Granitzer, M. (2010e). Efficient cross-domain classification of weblogs. *International Journal of Intelligent Computing Research*, 1(2).

[Lex et al., 2008] Lex, E., Seifert, C., Kienreich, W., and Granitzer, M. (2008). A generic framework for visualizing the news article domain and its application to real-world data. *Journal of Digital Information Management*, 6:434–441.

[Li et al., 2008] Li, Y., Dong, M., and Ma, Y. (2008). Feature selection for clustering with constraints using jensen-shannon divergence. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1 –4.

[Liberman and Lempel, 2009] Liberman, S. and Lempel, R. (2009). Approximately optimal facet selection. In *Proceedings of the CIKM'09*.

[Likert, 1932] Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 140:1–55.

[Lim et al., 2005a] Lim, C. S., Lee, K. J., and Kim, G. C. (2005a). Automatic genre detection of web documents. *Lecture Notes in Computer Science*, 3248:310–319.

[Lim et al., 2005b] Lim, C. S., Lee, K. J., and Kim, G. C. (2005b). Multiple sets of features for automatic genre classification of web documents. *Information Processing & Management*, 41:1263–1276.

[Losee, 1997] Losee, R. M. (1997). Comparing boolean and probabilistic information retrieval systems across queries and disciplines. *J. Am. Soc. Inf. Sci.*, 48:143–156.

[Macdonald et al., 2010a] Macdonald, C., Ounis, I., and Soboroff, I. (2010a). Overview of the trec-2009 blog track. In *Proceedings of TREC 2009*.

[Macdonald et al., 2010b] Macdonald, C., Santos, R. L., Ounis, I., and Soboroff, I. (2010b). Blog track research at trec. *SIGIR Forum*, 44:58–75.

[Magdy and Wanas, 2010] Magdy, A. and Wanas, N. (2010). Web-based statistical fact checking of textual documents. In *Proceedings of the SMUC'10*.

[Manning et al., 2008] Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

[Manning and Schuetze, 2003] Manning, C. D. and Schuetze, H. (2003). *Foundations of Statistical Natural Language Processing*. The MIT Press, 6 edition.

[Marchionini, 2006] Marchionini, G. (2006). Toward human-computer information retrieval. *Bulletin of American Society For Information Science And Technology*, 32(5):20.

[Marcus et al., 1994] Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1994). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

[Mathes, 2004] Mathes, A. (2004). Folksonomies - cooperative classification and communication through shared metadata. Available online. `http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html`.

[McCallum, 2005] McCallum, A. K. (2005). Information extraction: distilling structured data from unstructured text. *Queue*, 3(9):48–57.

[Microsoft Research, 2011] Microsoft Research (2011). Multiple kernel learning. http://research.microsoft.com/en-us/groups/vgv/, last accessed Jan 2011.

[Mishne and de Rijke, 2006] Mishne, G. and de Rijke, M. (2006). A study of blog search. In *ECIR 2006*.

[Mossberg, 2003] Mossberg, W. (2003). Mossberg's mailbox. *Wall Street Journal*, March 13.

[Muhr et al., 2010] Muhr, M., Kern, R., and Granitzer, M. (2010). Analysis of structural relationships for hierarchical cluster labeling. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 178–185, New York, NY, USA. ACM.

[Murakami et al., 2009] Murakami, K., Nichols, E., Matsuyoshi, S., Sumida, A., Masuda, S., Inui, K., and Matumoto, Y. (2009). Statement map: assisting information crediblity analysis by visualizing arguments. In *Proceedings of the 3rd workshop on Information credibility on the web*, WICOW '09, pages 43–50, New York, NY, USA. ACM.

[Murakami et al., 2010] Murakami, K., Nichols, E., Mizuno, J., Watanabe, Y., Masuda, S., Goto, H., Ohki, M., Sao, C., Matsuyoshi, S., Inui, K., and Matsumoto, Y. (2010). Statement map: reducing web information credibility noise through opinion classification. In *Proceedings of the fourth workshop on Analytics for noisy unstructured text data*, AND '10, pages 59–66, New York, NY, USA. ACM.

[Navigli, 2009] Navigli, R. (2009). Word Sense Disambiguation: A Survey. *ACM Computing Surveys (CSUR)*, 41(2):10.

[Nikulin, 2010] Nikulin, V. (2010). Web-mining with wilcoxon-based feature selection, ensembling and multiple binary classifiers. In *Proceedings of ECML/PKDD Discovery Challenge, Available online.*

[Nilsson, 1998] Nilsson, N. (1998). *Introduction to Machine Learning.* MIT Press (to apprear).

[O'Reilly, 2005] O'Reilly, T. (2005). O'Reilly Network: What Is Web 2.0. Available online. `http://www.oreillynet.com/lpt/a/6228`.

[Page et al., 1998] Page, L., Brin, S., Motwani, R., and Winograd, T. (1998). The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project.

[Pang and Lee, 2004] Pang, B. and Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the ACL.*

[Peters, 2009] Peters, I. (2009). *Folksonomies. Indexing and Retrieval in Web 2.0 (Knowledge and Information).* De Gruyter, 1 edition.

[Pew Research Center, 2011] Pew Research Center (2011). Internet gains on television as public's main news source. Available online. `http://people-press.org/report/689/,lastaccessedJan2011`.

[Pietramala et al., 2008] Pietramala, A., Policicchio, V., Rullo, P., and Sidhu, I. (2008). A genetic algorithm for text classification rule induction. In Daelemans, W., Goethals, B., and Morik, K., editors, *Machine Learning and Knowledge Discovery in Databases*, volume 5212 of *Lecture Notes in Computer Science*, pages 188–203. Springer Berlin / Heidelberg. 10.1007/978-3-540-87481-2$_1$3.

[Pujol et al., 2002] Pujol, M., Sangesa, R., and Delgado, J. (2002). Extracting reputation in multi agent systems by means of social network topology. In *Proc. of the first international joint conference on Autonomous agents and multiagent systems.*

[Qi and Davison, 2009] Qi, X. and Davison, B. D. (2009). Web page classification: Features and algorithms. *ACM Comput. Surv.*, 41:12:1–12:31.

[Qiu and Pang, 2008] Qiu, L.-Q. and Pang, B. (2008). Analysis of automated evaluation for multi-document summarization using content-based similarity. In *Proceedings of the Second International Conference on Digital Society*, ICDS '08, pages 60–63, Washington, DC, USA. IEEE Computer Society.

[Quinlan, 1993] Quinlan, J. R. (1993). *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[Raaijmakers and Kraaij, 2008] Raaijmakers, S. and Kraaij, W. (2008). A shallow approach to subjectivity classification. In *AAAI*.

[Ranganathan, 1933] Ranganathan, S. R. (1933). Colon classification. In *Madras Library Association*.

[Remus, 2011] Remus, R. (2011). Improving sentence-level subjectivity classication through readability measurement. In *Proceedings of the The 18th International Nordic Conference of Computational Linguistics (NODALIDA-2011)*. NEALT - to appear.

[Robertson and Gaizauskas, 1997] Robertson, A. and Gaizauskas, R. (1997). On the marriage of information retrieval and information extraction. In *Proceedings of BCS IRSG 19th Annual Colloquium on Information Retrieval Research*, pages 60–67, Aberdeen, Scotland.

[Rokach and Maimon, 2005] Rokach, L. and Maimon, O. (2005). Top-down induction of decision trees classifiers - a survey. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 35(4):476 – 487.

[Sabol et al., 2009] Sabol, V., Kienreich, W., Muhr, M., Klieber, W., and Granitzer, M. (2009). Visual knowledge discovery in dynamic enterprise text repositories. In *Proceedings of the 2009 13th International Conference Information Visualisation*, pages 361–368, Washington, DC, USA. IEEE Computer Society.

[Safavian and Landgrebe, 1991] Safavian, S. and Landgrebe, D. (1991). A survey of decision tree classifier methodology. *Systems, Man and Cybernetics, IEEE Transactions on*, 21(3):660 –674.

[Sakoe and Chiba, 1978] Sakoe, H. and Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*.

[Salton, 1971] Salton, G., editor (1971). *The SMART Retrieval System - Experiments in Automatic Document Processing*. Prentice Hall, Englewood, Cliffs, New Jersey.

[Salton and Buckley, 1988] Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24:513–523.

[Salton et al., 1975] Salton, G., Wong, A., and Yang, C. S. (1975). A vector space model for automatic indexing. *Commun. ACM*, 18:613–620.

[Salvadore and Chan, 2004] Salvadore, S. and Chan, P. (2004). FastDTW: Toward accurate dynamic time warping in linear time and space. In *3rd Workshop on Mining Temporal and Sequential Data*.

[Savage, 2010] Savage, N. (2010). New search challenges and opportunities. *Commun. ACM*, 53(1):27–28.

[Schapire, 1990] Schapire, R. E. (1990). The strength of weak learnability. *Mach. Learn.*, 5:197–227.

[Scharl and Tochtermann, 2007] Scharl, A. and Tochtermann, K. (2007). *The Geospatial Web: How Geobrowsers, Social Software and the Web 2.0 are Shaping the Network Society (Advanced Information and Knowledge Processing)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

[Schroeder et al., 1986] Schroeder, L. D., Sjoquist, D. L., and Stephan, P. E. (1986). *Understanding Regression Analysis: An Introductory Guide*. Sage Publications Inc.

[Sebastiani, 2002] Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Comput. Surv.*, pages 1–47.

[Segal and Akeley, 2006] Segal, M. and Akeley, K. (2006). The opengl graphics system: A specification, (version 2.1). Available online. `http://www.opengl.org/documentation/specs/version2.1/glspec21.pdf`. [Online; accessed 06-January-2011].

[Seifert et al., 2008] Seifert, C., Kump, B., Kienreich, W., Granitzer, G., and Granitzer, M. (2008). On the beauty and usability of tag clouds. In *Proceedings of the 2008 12th International Conference Information Visualisation*, pages 17–25, Washington, DC, USA. IEEE Computer Society.

[Seifert and Lex, 2009] Seifert, C. and Lex, E. (2009). A novel visualization approach for data-mining-related classification. *Information Visualisation, International Conference on*, 0:490–495.

[Sen et al., 2006] Sen, S., Lam, S. K., Rashid, A. M., Cosley, D., Frankowski, D., Osterhouse, J., Harper, F. M., and Riedl, J. (2006). tagging, communities, vocabulary, evolution. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*, CSCW '06, pages 181–190, New York, NY, USA. ACM.

[Serdyukov, 2010] Serdyukov, P. (2010). Faceted search tutorial at www 2010. Available online. `http://www.slideboom.com/presentations/163195/Faceted-Search-Tutorial-at-WWW-2010`.

[Singhal, 2001] Singhal, A. (2001). Modern Information Retrieval: A Brief Overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 24(4):35–42.

[Smyth et al., 2011] Smyth, B., Coyle, M., and Briggs, P. (2011). Communities, collaboration, and recommender systems in personalized web search. In Ricci, F., Rokach, L., Shapira, B., and Kantor, P. B., editors, *Recommender Systems Handbook*, pages 579–614. Springer.

[Sokolov et al., 2010] Sokolov, A., Urvoy, T., Denoyer, L., and Ricard, O. (2010). Madspam consortium at the ecml/pkdd discovery challenge 2010. In *Proceedings of ECML/PKDD Discovery Challenge, Available online.*

[Sokolova et al., 2006] Sokolova, M., Japkowicz, N., and Szpakowicz, S. (2006). Beyond Accuracy, F-score and ROC: A family of discriminant measures for performance evaluation. In *Advances in Artificial Intelligence (AI2006)*, LNAI 4304, pages 1015–1021, Berlin/Heidelberg. Springer.

[Spoustová et al., 2009] Spoustová, D., Hajič, J., Raab, J., and Spousta, M. (2009). Semi-supervised training for the averaged perceptron pos tagger. In *Proceedings*

of the 12th Conference of the European Chapter of the Association for Computational Linguistics, EACL '09, pages 763–771, Morristown, NJ, USA. Association for Computational Linguistics.

[Stamatatos et al., 2000] Stamatatos, E., Fakotakis, and Kokkinakis, G. (2000). Text genre detection using common word frequencies. In *Proceedings of the 18th conference on Computational linguistics*, pages 808–814.

[Steele, 2009] Steele, B. M. (2009). Exact bootstrap k-nearest neighbor learners. *Mach. Learn.*, 74:235–255.

[Stein et al., 2010] Stein, B., Meyer zu Eißen, S., and Lipka, N. (2010). *Genres on the Web*, volume 42 of *Text, Speech and Language Technology*, chapter Web Genre Analysis: Use Cases, Retrieval Models, and Implementation Issues, pages 167–190. Springer, Berlin Heidelberg New York.

[Strapparava and Mihalcea, 2007] Strapparava, C. and Mihalcea, R. (2007). Semeval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74, Prague, Czech Republic. Association for Computational Linguistics.

[Sun et al., 2007] Sun, A., Suryanto, M. A., and Liu, Y. (2007). Blog classification using tags: An empirical study. *Lecture Notes in Computer Science*, pages 307–316.

[Tan et al., 2006] Tan, P.-N., Steinbach, M., and Kumar, V. (2006). *Introduction to Data Mining*. Pearson Education.

[Taylor, 1992] Taylor, A. G. (1992). *Wynar's Introduction to Cataloging and Classification*. Libraries Unlimited, 8th edition edition.

[Taylor, 2003] Taylor, C. (2003). An introduction to metadata. *Technology*, pages 1–5.

[Toffler, 1984] Toffler, A. (1984). *Future Shock*. Bantam, reissue edition.

[Tunkelang, 2006] Tunkelang, D. (2006). Dynamic Category Sets: An Approach for Faceted Search. In *Proceedings of the SIGIR '06 Workshop on Faceted Search Conference*.

[Tunkelang, 2009] Tunkelang, D. (2009). *Faceted Search*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers.

[Tunkelang, 2010] Tunkelang, D. (2010). Linkedin signal = exploratory search for twitter. http://thenoisychannel.com/2010/10/02/linkedin-signal-exploratory-search-for-twitter/comment-7008.

[Vapnik, 1995] Vapnik, V. N. (1995). *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA.

[Voorhees, 2005] Voorhees, E. M. (2005). Trec: Improving information access through evaluation. *Bulletin of the American Society for Information Science and Technology*, 32(1):16–21.

[Wai-chee et al., 2005] Wai-chee, Ada, F., Keogh, E., Lau, L. Y. H., and Ratanamahatana, C. A. (2005). Scaling and time warping in time series querying. *VLDB*.

[Wallach, 2006] Wallach, H. M. (2006). Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd international conference on Machine learning*, ICML '06, pages 977–984, New York, NY, USA. ACM.

[Wang and Strong, 1996] Wang, R. Y. and Strong, D. M. (1996). Beyond accuracy: what data quality means to data consumers. *J. Manage. Inf. Syst.*, 12:5–33.

[White and Roth, 2009] White, R. W. and Roth, R. A. (2009). Exploratory Search: Beyond the Query-Response Paradigm. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 1(1):1–98.

[Wiebe and Bruce, 1999] Wiebe, J. and Bruce, R. O. T. (1999). Development and use of a gold standard data set for subjectivity classifications. In *Proceedings of ACL-99*.

[Wiebe and Mihalcea, 2006] Wiebe, J. and Mihalcea, R. (2006). Word sense and subjectivity. In *Joint conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics. (COLING-ACL 2006)*.

[Wilson et al., 2005] Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 347–354, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Wilson, 2008] Wilson, T. A. (2008). *Fine-grained Subjectivity and Sentiment Analysis: Recognizing the Intensity, Polarity, and Attitudes of Private States*. PhD thesis, University of Pittsburgh.

[Witten and Frank, 2005] Witten, I. H. and Frank, E. (2005). *Data mining : practical machine learning tools and techniques*. Elsevier, Morgan Kaufman, Amsterdam [u.a.], 2. ed. edition.

[Wolpert and Macready, 1997] Wolpert, D. and Macready, W. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*.

[Wu et al., 2004] Wu, T.-F., Lin, C.-J., and Weng, R. C. (2004). Probability estimates for multi-class classification by pairwise coupling. *J. Mach. Learn. Res.*, 5:975–1005.

[Xue et al., 2008] Xue, G.-R., Dai, W., Yang, Q., and Yu, Y. (2008). Topic-bridged plsa for cross-domain text classification. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 627–634, New York, NY, USA. ACM.

[Yang et al., 2007] Yang, C., Lin, K. H.-Y., and Chen, H.-H. (2007). Emotion classification using web blog corpora. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*.

[Yoo, 2011] Yoo, Dong Kyoon, V. M. A. R.-N. T. (2011). Knowledge quality: antecedents and consequence in project teams. *Journal of Knowledge Management*, pages 329–343.

[Zelevinsky, 2010] Zelevinsky, V. (2010). Breaking down the assumptions of faceted search. In *HCIR'10*.

[Zhang, 2000] Zhang, G. (2000). Neural networks for classification: a survey. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 30(4):451 –462.

[Ziegler and Lausen, 2005] Ziegler, C.-N. and Lausen, G. (2005). Propagation models for trust and distrust in social networks. *Information Systems Frontiers*, pages 337–357.