VANESSA FEICHTNER, BSc

Information Quality using Social Network Analysis in User Generated Content

Master's Thesis

to achieve the university degree of

Diplom-Ingenieurin

Master's degree programme: Software Development and Business Management

submitted to

Graz University of Technology

Supervisor
Ass.Prof. Dipl.-Ing. Dr.techn. Lex Elisabeth

Institute for Knowledge Technologies
Head: Univ.-Prof. Dipl.Ing. Dr.techn. Lindstaedt Stefanie

Faculty of Computer Science

Graz, March 2016

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.

Graz, _____ _____

Date Signature

Eidesstattliche Erklärung[1]

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 entnommenen Stellen als solche kenntlich gemacht habe.

Graz, am _____ _____

Datum Unterschrift

[1] Beschluss der Curricula-Kommission für Bachelor-, Master- und Diplomstudien vom 10.11.2008; Genehmigung des Senates am 1.12.2008

Abstract

Wikipedia is a popular online encyclopaedia which anyone can edit. The access to its content is free and heavily used. Many people use it to get a first impression on a topic before going deeper into theory. Due to the encyclopaedic nature of Wikipedia, the articles should be of high quality and standardized in terms of structure. In order to classify articles Wikipedia introduced the featured article tag which indicates a high quality article. Articles have to be nominated by users and are tagged featured after a time consuming reviewing process. As this thesis tries to analyse the quality of articles based on their network properties, the Wikipedia dumps of the English and Norwegian Wikipedia are downloaded and then preprocessed. From these data sets four different networks are constructed, the article-, user-, collaboration- and a two-mode network. Afterwards several metrics from Social Network Analysis are calculated. Depending on the type of network, these metrics are article length, local or extended clustering coefficient, average path length and betweenness. These metrics are successfully used to find differences in the properties between featured and non-featured articles. This thesis shows also that the conclusions of the underlying paper by Ingawale et al., that structural holes indicate good quality articles, is true. For this reason the clustering coefficient together with betweenness is used to find featured articles at structural holes. Afterwards with the metrics mentioned before, a naive Bayes 10-fold cross validation is performed on the networks and compared to different combinations of metrics. The results for two-mode network show that this type of network is not able to automatically classify featured articles. The classification for the article network produces only weak results in correctly classifying articles.

Keywords. Wikipedia, Social Network Analysis, Classification

Kurzfassung

Wikipedia ist eine beliebte online Enzyklopädie, die jeder verändern kann. Der Zugriff auf den Inhalt der Artikel ist frei und sehr beliebt. Viele Leute nutzen Wikipedia um einen ersten Eindruck über ein Thema zu erhalten, bevor sie sich näher damit auseinandersetzen. Aufgrund des enzyklopädischem Aufbau der Wikipedia, sollten die Artikel von guter Qualität sein, sowie im Aufbau des Inhalts standardisiert. Um qualitativ hochwertige Artikel zu markieren, hat Wikipedia spezielle Kennzeichen eingeführt. Artikel werden zuerst nominiert und nach einer zeitaufwändigen Qualitätsüberprüfung als *featured* Artikel markiert. Da diese Arbeit versucht die Qualität der Artikel aufgrund ihrer Netzwerkeigenschaften zu analysieren, wurden der englische und norwegische Wikipedia Datenbestand heruntergeladen und verarbeitet. Aus diesem Datenbestand wurden vier verschiedene Netzwerke generiert, das Artikel-, User-, Interaktions- und das Two-Mode Netzwerk. Danach wurden einige Metriken aus dem Bereich Social Network Analysis berechnet. Je nach Netzwerkart sind das Artikellänge, local und Extended Clustering Coefficient, durchschnittliche Weglänge und der Betweenness Wert. Diese Metriken wurden erfolgreich verwendet um Unterschiede zwischen featured und nicht featured Artikel zu finden. Die Arbeit zeigt auch, dass die Erkenntnisse der zugrundeliegenden Arbeit von Ingawale et al., nämlich, dass structural holes auf Artikel mit guter Qualität schließen lassen, auf diesen Datenbestand zutrifft. Danach wurden die zuvor berechneten Metriken verwendet um eine naive Bayes 10-fold cross Validierung, mit Kombinationen dieser Metriken durchzuführen. Die Ergebnisse des Two-Mode Netzwerks lässt sich nicht für eine automatische Klassifizierung zwischen featured und nicht featured Artikel verwenden. Auch die Klassifizierung mit dem Artikel Netzwerk liefert nur eine unzureichende Genauigkeit.

Acknowledgements

I would like to thank my supervisor Ass.-Prof. Dipl.-Ing.Dr.techn. Elisabeth Lex for giving me the opportunity to work on this thesis and supporting me during especially in the last months.
Thanks also goes to my parents, who always believed in me and gave me the possibility to study whatever I wanted.
I would also like to thank all my friends for always helping me during my studies and having fun during long lessons, especially Katharina and Christoph, who always had an open ear for my troubles.
A great thanks goes to my boyfriend Peter who always supported me and gave me security in what I did.

Contents

Contents

1. Introduction and Research Questions

Wikipedia is a fast growing online encyclopaedia which anyone can edit. It was funded in 2001 and is currently among the top ten most visited websites ("Analysing Web-Traffic"). Since Wikipedia is free and access to the World Wide Web has become natural, it is one of the first websites to be consulted, when looking for information. Therefore the content should be reliable and trustworthy. Compared to traditional encyclopaedias Wikipedia is solely written by unpaid volunteers. Therefore there are many advantages and disadvantages due to the fact that anyone can edit Wikipedia articles. A disadvantage is that even non-specialists on a topic can edit articles, but the advantage is, that the wisdom of the crowd often detects errors and correct them quite fast. Wikipedia introduced different labels for articles where readers can see, whether this article is of good quality. The label, indicating the best quality an article can have, is the *featured articles* label. Articles with this label had to undergo many reviews and discussions between contributors, until they agreed, that the information in this article was correct and considered to be of good quality. Since this is a tedious and slow process, there are only very few *featured articles* in Wikipedia. But this few articles seem to have some special properties, that non featured articles do not have. Based on the paper by Ingawale et al., which proposes to identify featured articles using network analysis, this thesis investigates differences in article contents, edit statistics and network structure between featured and non-featured articles. Based on our findings we also want to automatically classify featured articles. This might help the community of Wikipedia to faster label articles as featured.
For this work the Norwegian Wikipedia is used due to its smaller size compared to for example the English Wikipedia. In some parts also the English

Wikipedia is used, but since it is ten times larger than the Norwegian data set, it was only used for the classification part. This thesis will not only look at the article contents and edit statistics itself, but also on the four different types of networks generated from the dataset. These four networks all have different interpretations and properties. For the English Wikipedia only the article network was generated, whereas for the Norwegian Wikipedia also the user-, collaboration- and two-mode network are generated (see section Social Network Analysis).
This thesis tackles four Research Questions, which are answered in chapter Experiments & Results.

1.1. Research Questions

Research Question 1: Do featured articles have on average different article or network related properties than non-featured articles?

Featured articles represent high quality articles in Wikipedia and are quite rare due to its tedious selection process. Therefore featured articles might have different properties both in content statistics and network structure. In the content statistics the article length will be analysed. In the different networks the metrics clustering coefficient, betweenness, redundancy and average path length will be analysed in order to see differences in the properties of featured and non-featured articles.

Research Question 2: Do featured articles build a bridge between categories? Do they lie at nodes spanning structural nodes?

As featured articles try to cover their topic from all angles, they touch and reference many related topics. Therefore it is possible that these connections build bridges between different categories of articles. This research question is also stated in the underlying paper of (Ingawale et al., 2013) and summarized in chapter Methodology.

Research Question 3: Can articles be classified into featured and non-featured articles with given properties? And which properties are most useful in order to classify featured articles?

Since the selection process of featured articles is long and time consuming it might be very useful to be able to classify articles automatically. Articles that are classified as featured articles but are not labelled as such yet might go on a suggestion list for the contributors to look at. This would speed up the selection process and save some time.

Research Question 4: Can the analysis of a one-mode network be extended to a two-mode network? Does the two-mode network show similar differences in the properties for featured and non-featured articles and can featured articles automatically be classified?

This thesis analyses four different networks, from which three are one-mode networks and one is a two-mode network with articles and user as nodes. Ingawale et al. use a projection of a two-mode network, which becomes quite dense. Therefore the original two-mode network is analysed in this thesis, in order to find similar properties as for the one-mode network.

1.2. Overview

This thesis is structured into six chapters:
At first Introduction and Research Questions gives a quick overview of this thesis and states the research questions. It also gives a the motivation for this thesis. The next chapter introduces to Wikipedia and the history and growth of the English and Norwegian Wikipedia. Afterwards Information Quality in Wikipedia discusses three different types on finding good quality articles in Wikipedia together with their problems.
Methodology at first introduces the underlying paper by Ingawale et al. and then gives an introduction to Social Network Analysis and the four different networks that are generated. After that, different network metrics, the definition of structural holes and the classification process are introduced.

Chapter Experiments & Results contains all experiments made on the different networks and its results. In this chapter the research questions will be answered.
And lastly Conclusion summarizes the findings in this thesis and outlines further possible investigations.

2. Wikipedia

Wikipedia is a free online encyclopaedia which anyone can edit. This chapter focuses on the history and growth of Wikipedia. At first the history, features and mark-up of Wikipedia are introduced. Then the current sizes of the currently four biggest Wikipedias is shown. Afterwards the growth of the English and Norwegian Wikipedia is presented.

2.1. History of Wikipedia and Mediawiki

In March 2000 the online encyclopaedia *Nupedia* was introduced by Jimmy Wales and Larry Sanger. It had a sophisticated peer-review process, in order to have good quality articles. The project was financed by "Bomis", an online host for forums for sport, women and science-fiction. Due to its time expensive peer-review process, Nupedia had 25 finished articles at the end of September 2003, and 74 were still in progress. By then the project was cancelled because of inefficiency and the new project - Wikipedia - was forced instead.
In 2001 Jimmy Wales introduced *Wikipedia* as a prestage for Nupedia. The goal was that people write articles that other people can correct. After this content creation process, the peer-review for the Nupedia would have been much faster and less time consuming. But Wikipedia was such a success that Nupedia was quit permanently and Wikipedia was the main system that was used then (*Nupedia*).
The software behind Wikipedia is developed by the *Wikimedia Foundation* and is called *Mediawiki*. Mediawiki is written in PHP and uses MySQL as its database backend. It also supports Oracle, SQLite, PostgresSQL and MariaDB. The development of Mediawiki started in 2002. In July 2003 the name of the software - Mediawiki - was chosen. It is a open source software,

and more than 60 programmers and contributors worked on Mediawiki in 2005. According to Wikimedia, Mediawiki is a very successful open source software and is used by numerous companies and organisations (*Nupedia*).

2.2. Features of Mediawiki

Mediawiki supports collaborative editing and creation of user generated content. Articles can be connected by using links and therefore Wikipedia forms a graph structure. Mediawiki is web-based and uses a special Mediawiki mark-up for the notation of articles which allows the user to easily format the content of articles (see section Mark-up). Mediawiki also offers numerous features in order to categorize or to keep track of changes of articles. The most important features for this thesis are explained now (*MediaWiki*):

- **Categorization and Namespaces:** Mediawiki supports several methods to differentiate between articles. Articles can belong to a category (but do not need to), and a namespace (mandatory). Namespaces categorize different types of articles. The most important one is namespace 0, because it contains the general articles, in which the typical encyclopaedic knowledge is stored. Other namespaces are for example *Talk* pages (where the discussion for an article takes place) or *User Profile* sites. All in all there are 28 different namespaces to which an article can belong, but it can only belong to one namespace at all.
- **Versioning:** All changes on articles and its associated mediafiles, like images or videos, which are made by users or automated bots, are recorded in the revisions. Either the user name and user id or the IP-address is stored together with the time of modification. Therefore it is possible to reconstruct the revision of an article at any possible time and to trace back changes.
- **Templates:** Text blocks that are often used can be saved as a template. For example the boiler plate text to identify a featured article consists of a template, which is used in every featured article. This is useful as it makes articles more standardized, which is a difficult task in a collaboration process in general.

- **Interwiki-Links** Mediawiki supports a special type of link, which can be used to connect articles of other Mediawiki based projects. In the case of Wikipedia these links can be used to connect to Wikimedia Commons content.
- **Interlanguage-Links:** Interlanguage links are used to connect instances of the same article in other languages of the Wikipedia.
- **Most recent changes:** Mediawiki displays the most recent changes of an article. The revisions of the article can also be compared. Additionally Mediawiki provides RSS- and Atom-Feed in order to keep track of changes.
- **Right and Role Management:** Mediawiki allows custom definition of user roles. Each role can have several rights associated. Administrators can deny write permissions for a given article to certain users. In the case of Wikipedia everybody - registered or unregistered - is allowed to make changes for almost every article in namespace 0. However, there are access limitations for articles that are prone to vandalism, for example there is a semi-protection for unregistered users for the article "Autism" in the English Wikipedia (*Autism*).
- **Full-text search:** Mediawiki uses Apache Lucene to provide full-text search capabilities.

2.3. Mark-up

The mark-up language of Mediawiki is very extensive. Therefore this section will only outline some mark-up features of Mediawiki, which are relevant for the data extraction process which is explained in more detail in section Processing Data Set.

Mark-up for Flagging Articles As Wikipedia articles do not have any flag attributes attached to them in the database, templates are used to attach attributes to an article.
Curly braces {{ signify the use of a named template for example the presence of a *featured article*, a *disambiguation site* or a *redirect*. The data extraction process is using the disambiguation and the redirect template in order to

detect a given article as a redirect or disambiguation page and to remove it from the cleaned data set.

For example a featured article is marked as such, by the use of {{`featuredarticle`}} in its content. This tag is not visible as text on the page itself, but the mediawiki software detects this and draws a star at the right upper corner of the article (*Wikipedia:Featured articles*).

Mark-up for Formatting and Styling As Wikipedia articles are intended to be human readable, it is required to have formatting and styling features. Mediawiki provides a markup to highlight text, to structure the article in sections and headlines, and to style the text flow or to embed media files. The following example shows a large headline for a section and a smaller one for a subsection:

```
==Headline1==
===Headline2===
```

As the styling markup does have no significance for the parser it is not further explained. For more information on this topic the reader is asked to have a look at *Help:Wiki markup*.

Link mark-up Wikipedia offers several kinds of links. This thesis only looks at links between articles within the Wikipedia. Therefore this section only outlines the mark-up for link with the target inside the same Wikipedia. This mark-up creates a hyperlink to the article named in the braces.

```
[[LinkTarget]] or [[LinkTarget|Link Display Name]]
```

The name of the link is case sensitive, as there may be two articles which only differ in the casing of their title. If the link tag contains a "|" character, the tag contains an additional description, which will be displayed as the link text instead. A link is displayed as blue color in the Wikipedia and a click on it leads to the corresponding Wikipedia article.

2.4. Sizes of Different Wikipedias

At the time of December 2015 there were 280 active Wikipedias hosted by the Wikimedia foundation (*List of Wikipedias*). The largest and most active Wikipedias are the English, Swedish and the German Wikipedias. According to Alexa all Wikipedias combined are currently in the top 10 of the most visited websites of the internet ("Analysing Web-Traffic").
The sizes of the different Wikipedias are shown now in order to give an impression of its dimensions. The Norwegian Wikipedia is also included to this table 2.1 because of its relevance to this master thesis.

Language	Articles	Edits	Admins	Users	Active Users
English	5,035,665	804,867,810	1,332	27,035,139	125,367
Swedish	2,279,871	31,662,170	73	470,385	2,758
German	1,887,197	154,333,342	246	2,314,750	19,031
Norwegian	427,469	15,016,191	49	346,716	1,549

Table 2.1.: Comparison of the different Wikipedias, (*List of Wikipedias*)

When referring to articles only the content from namespace 0 is meant. As mentioned before the English Wikipedia is the biggest one with more than 5 million articles. The Swedish Wikipedia has more articles, but the German Wikipedia has more active users and edits. It might be interesting if the quality of articles in the German Wikipeda is therefore better, or if the Swedish Wikipedia writes better quality articles in the first place. This might be discussed in a different thesis or paper.

2.5. Growth and Size of English Wikipedia

When looking at the growth of Wikipedia, one has to consider several aspects. Growth of Wikipedia can be measured by different means:

- Rate of new articles that are added
- Rate of articles that are modified
- Length of articles

The growth of articles in the English Wikipedia initially started with what seemed to be an exponential growth rate. In 2007 this rate peaked and was slowed down (see figure 2.1). Several factors seem to be the reason for this slowing of growth, including protectionism, user blocks and vandalism (Suh et al., 2009). But the slowing growth might also be explained as the number of new editors is slowing down, since the chance is low to make new articles, as the coverage of articles in Wikipedia is broad (Suh et al., 2009; Royal and Kapila, 2008).
Over the last years Wikipedia received more and more new articles and therefore also the number of links grew (Buriol et al., 2006). In the time interval of 2003 to 2006 the number of articles grew at a 6.1% rate a month and the number of unique editors grew at a 13% rate (Buriol et al., 2006). The growth of new editors in the Norwegian Wikipedia will be described in section Growth and Size of Norwegian Wikipedia.
The article growth in the English Wikipedia can be described with a *Gompertz Curve* (*Wikipedia:Modelling Wikipedia's growth*):

$$\begin{aligned} y &= ae^{be^{ct}} \\ a &= 4,378,449 \\ b &= -15.42677 \\ c &= -0.384124 \\ t &= 10 \end{aligned}$$

The real number of articles grew more than the Gompertz curve would estimate. By the end of 2015 there were more than 5,000,000 articles in the English Wikipedia, were the Gompertz estimation suggested only about 4,250,000. Since the parameters for the Gompertz curve were chosen in 2000 to estimate a timespan of 10 years, there should be some new evaluations done for new approximations. The growth of new articles in the Norwegian Wikipedia is described in chapter Growth and Size of Norwegian Wikipedia.

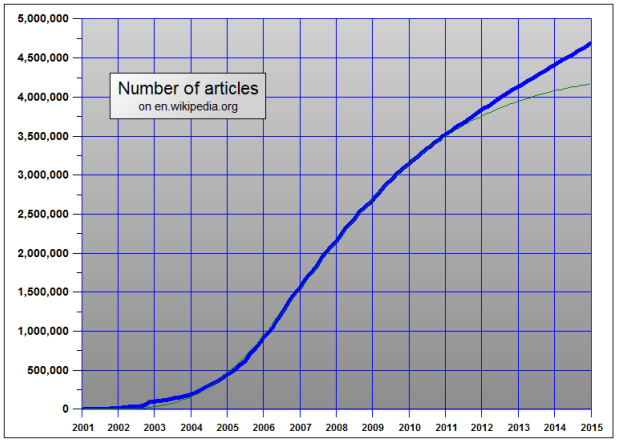


Figure 2.1.: Growth of Wikipedia compared to the Gompertz Curve *Number of articles on en.wikipedia.org and Gompertz extrapolation*

2.6. Growth and Size of Norwegian Wikipedia

The Norwegian Wikipedia was founded on November 26th, 2001 and was the sixteenth language edition of Wikipedia (*Wikipedia*). Since then it grew exponentially and is now the twentieth biggest Wikipedia.

	Data	Yearly Change	Monthly Change
Page Views per Month	30,623,246	-	-
Article Count	425,490	+6%	+1%
New Articles per Day	80	-	-
Edits per Month	59,940	+57%	−80%
Active Editors	390	+3%	−11%
Very Active Editors	71	+25%	−7%
New Editors	55	−21%	−27%
Speakers	4,700,000	-	-
Editors per Million Speakers	83	-	-

Table 2.2.: Statistics of Norwegian Wikipedia, November 2015 (*Wikimedia Statistics*)

The statistics in table 2.2 only counts articles and edits on articles that are

made in namespace 0, which means that these articles are the encyclopaedic content of Wikipedia. Editors that are registered and have more than five edits per month are considered *Active Editors*, whereas editors that are registered and have more than 100 edits per month are considered *Very Active Editors*. *New Editors* are editors that are registered and made their 10th edit in this month. *Speakers* are native and secondary language speakers and *Editors per Million Speakers* defines the participation rate (*Wikimedia Statistics*).
The number of articles is steadily growing since 2004. Compared to the English Wikipedia the Norwegian Wikipedia is much smaller and younger and therefore the growth rate is not declining yet (see figure 2.2).

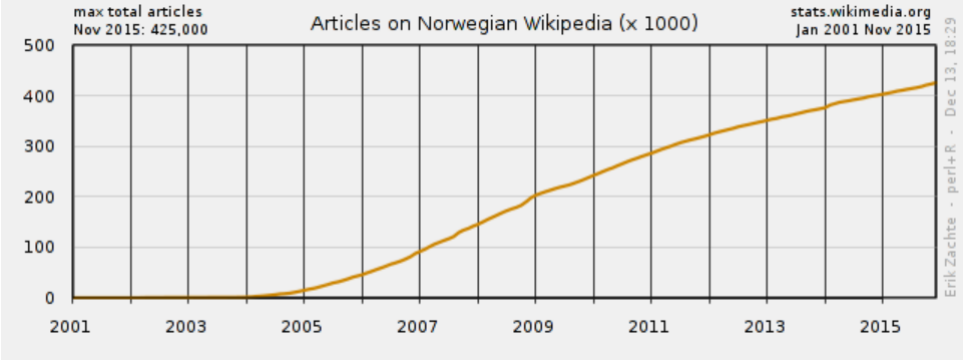


Figure 2.2.: Growth of Articles in Norwegian Wikipedia (*Wikimedia Statistics*)

On average there are 80 new articles per day which are generated by registered users (reg), anonymous users (anon) or bots. With theses three types of contributors there is a total of 425,490 articles in November 2015 (see figure 2.3).

All these articles grow only with the edits that contributors make. There are new edits but there may also be reverts due to vandalism or wrong information (see figures 2.4 and 2.5).

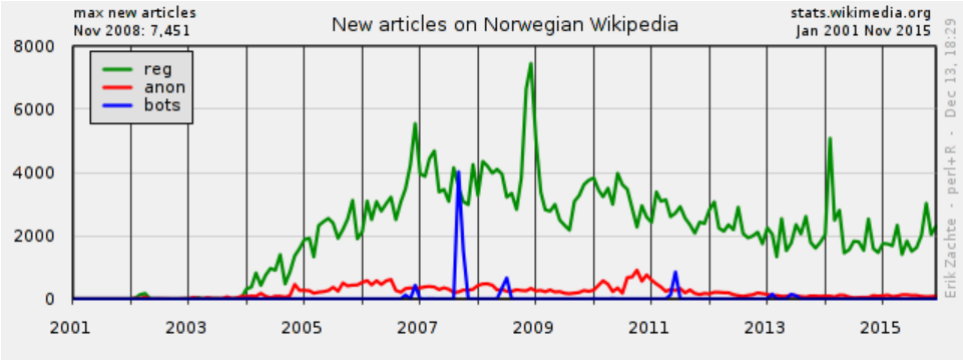


Figure 2.3.: Average new articles for Norwegian Wikipedia (*Wikimedia Statistics*)

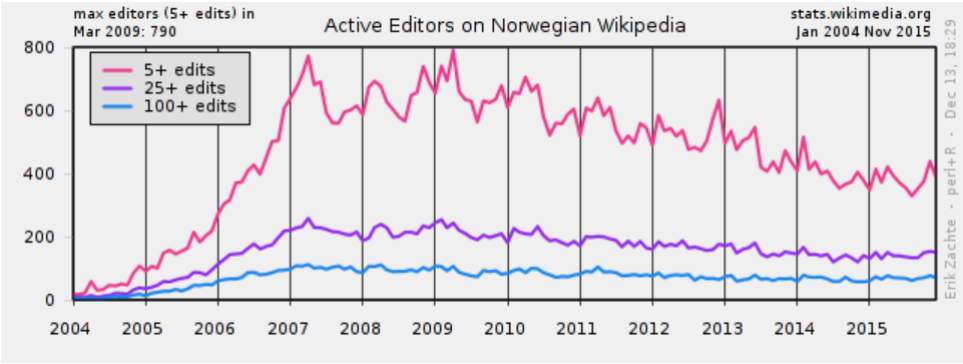


Figure 2.4.: Active Editors on Norwegian Wikipedia (*Wikimedia Statistics*)

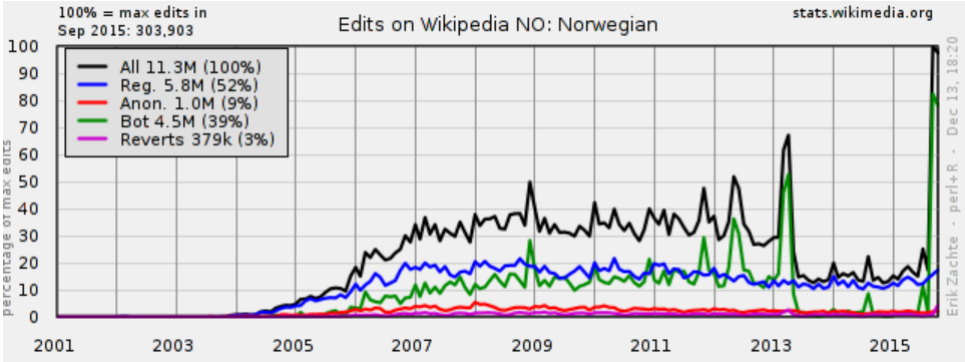


Figure 2.5.: Edits on Norwegian Wikipedia from 2001 to 2015 (*Wikimedia Statistics*)

2.7. Quantity versus Quality of Wikipedia Articles

Not only the size in terms of the number of articles is relevant to the user, but also the information quality of the articles. Wikipedia uses a classification scheme, in order to classify the different types of quality. Articles can be *Featured*-class, *A*-class, *Good Article*-class, *B*-class, *C*-class, *Start*-class or *Stub*-class articles, from *Featured*-class being the best quality to *Stub*-class being the first draft version of an article.
The English Wikipedia by December, 2015 consists of the following articles:

- Total Articles: 5,035,665
- Good Articles: 23,133
- Featured Articles: 4,699

This tables shows that only a very small number of articles actually meet the quality requirements to be classified as Good or even Featured Article in the English Wikipedia. The Norwegian Wikipedia until November 2015 has 254 Featured Articles from 162,853 total articles (*Utmerkede artikler*).
This classification is the result of a community process. Articles can be nominated for a specific class by contributors and will then be promoted to be for example featured article state. As this process cannot be done without manual human intervention, it is very time consuming, and therefore the chance that articles are simply not promoted because they were not found by the review community is high. Wikipedia uses templates in order to tag an article as a specific class-article. Since content creation is a community process and the use of templates is not mandatory, the chances are high that not all articles have a corresponding template included in its text. For this reason this thesis tries to find metrics to use a classifier, which can automatically label article as featured or non-featured.

3. Information Quality in Wikipedia

Wikipedia articles are user generated content, which by nature makes it hard to standardize all articles and to have them on the same quality level. Therefore Wikipedia introduced different kinds of quality levels. It starts with a *stub*, goes on to *good article* and in the best case an article becomes a *featured article*. However, for an article to be considered as a *featured article* it must full fill certain requirements. Its textual content has to be *well written*, defined by Wikipedia (*Wikipedia:Featured article criteria*) by the following criteria:

1. **well-written:** Its prose has to be written engagingly and it has to be of professional quality.
2. **comprehensive:** It completely covers all facets of the given topic.
3. **well-researched:** The article and all its claims and facts are documented by reliable high quality sources. It covers relevant sources, and uses in-line citation to make the statements verifiable.
4. **neutral:** It represents a neutral and fair point of view. It has to be unbiased.
5. **stable:** Changes to the article occur seldom, there are no edit wars in progress which could signify a dispute about some parts of the article. Changes may only be made according to the featured article process.

In regard to styling and formatting, the following guidelines are of importance for featured articles:

1. **A lead paragraph is provided:** A precise summary of the article is provided, which gives a quick overview and prepares for more detailed sections of the article.

2. **Appropriate structure:** The article provides a table of contents, which gives an overview of the hierarchical division of the content into sections and subsections.
3. **Consistent Citations:** Articles should use one of the allowed citation styles (e.g. Harvard or footnotes) for in line citations and provide a complete bibliography for all its citations.
4. **Media:** The article feature images and other media to illustrate the topic. Each media is described by a caption text and has a acceptable copyright status. All images in the article follow Wikipedias image use policy. If non-free images are used, then they have to follow certain rules and must be labelled and marked as such.
5. **Length:** The entire text of the article has to stay on the topic and must not deviate.

Wikipedia tries to standardize the quality of articles with these guidelines. How information quality of an article can be captured is discussed in the following sections.

3.1. Information Quality in Wikipedia by Collaboration Structure

A collaboration network in the context of Wikipedia means that the *graph nodes* consist of *articles* and the *edges* connecting them of *users* that contributed to the same article. This network structure is a *scale-free network* which means that the distribution follows a power-law (Stvilia et al., 2005) and therefore exhibits the small-world property (Goh et al., 2002). However, it is also interesting because scale-free networks also exhibits growth and preferential attachment (Stvilia et al., 2005).
Most studies on the topic of information quality in Wikipedia by collaboration structure define an *edge* if two editors collaborated on the same article. But not every edit to an article has the same importance, for example adding a simple comma to a sentence is not as important as a whole new paragraph. In order to diminish the number of edges in the collaboration network (as otherwise the network can become quite dense (Laniado and Tasso, 2011)), one can look for the main user, for whom most of his content

was accepted by the other users (Laniado and Tasso, 2011). For this reason an algorithm with three steps is introduced (Laniado and Tasso, 2011):

- Calculate score, where the contribution of each contributor to an article is measured with *edit longevity*. This value measures not only the count of changed text but also its acceptance over time (Adler et al., 2008).
- Identify main contributors to an article. For this reason only users with a nickname are taken and then those are chosen whose edit longevity value is above a given threshold θ.
- Construct collaboration network where two articles are connected if they were edited by the same main authors.

With this algorithm Laniado and Tasso got an affiliation network with not more than 20 contributors to an article. They also verify the small-world and the scale-free network property and found out that there is often a lead author for an article.
Other researchers try to find good articles by saying that featured articles are often edited by good authors and vice versa (Hu et al., 2007). The problem with this approach is that authors are often experts in only a few fields (Qin and Cunningham, 2012). For this reason Qin and Cunningham not only use *edit longevity*, but also *author centrality*. For calculating *author centrality* the *degree centrality*, *betweenness centrality* and the *eigenvector centrality* are combined. Together with these two measures - *edit longevity* and *author centrality* - three different models are described by Qin and Cunningham:

- **Contribution-based model:** In this model the article quality raises when the edit longevity raises for every edit to this article.

$$Longevity_QScore(p) = \sum_{a \in A_p} contr(a, p) \tag{3.1}$$

  where a are the contributors to article p.
- **Centrality-based model:** If the contributor of an article is relatively central in the Wikipedia talk network or co-author networks, then the article will also be of good quality.

$$Cen_QScore(p) = \sum_{a \in A_p} centrality(a) \tag{3.2}$$

- **Combination of edit contribution and contributor authoritativeness:** For this model the two values for contribution and centrality are first normalized and then multiplied.

$$AuthorContr(a,p) = contr(a,p) * centrality(a) \quad (3.3)$$

$$Com_QScore(p) = \sum_{a \in A_p} AuthorContr(a,p) \quad (3.4)$$

After evaluating the results Qin and Cunningham state that it is indeed useful to take the combination of these two metrics. They also conclude that articles that received the main contribution by established authors seem to be of better quality, and that these articles show more collaboration between authors (Qin and Cunningham, 2012). With this method, experienced users might be asked to work on low-quality articles to improve its quality. However, with this method there are still some issues to discuss. For example the author contribution can be wrongly high if the user reverted a malicious edit, where he simply reverts it to the previous state (Qin and Cunningham, 2012).

3.2. Information Quality in Wikipedia by Network Structure

The Wikipedia network structure can be analysed by looking at the position of an article in the network. The network is built with the *articles* as *nodes* and with the *links* from one article to another as *edges*. A good quality article is not only one that was written or edited by users, that often contribute to good articles, but also one that has many citations to other articles, which may be an indication for its relevance (Hasan Dalip et al., 2009). There are several methods to calculate the importance of an article in the Wikipedia network (Hasan Dalip et al., 2009):

- **In Degree:** Counts the number of incoming links to this article. If an article has ahigh in degree value, it is an evidence for a high popularity. However, articles with low in degree may also be important (Kamps and Koolen, 2008).

- **Out Degree:** Counts the number of outgoing links to already existing articles.
- **Link Count:** Counts the number of outgoing links to all articles (even to those, that still have to be written).
- **Clustering Coefficient:** A value which represents how connected the linked neighbours of an article are.
- **PageRank:** This coefficient is an advancement of the eigenvector which give information about the popularity of an article (Brin and Page, 2012).

The list contains only a few of various measures of article quality, but it should give a short introduction to the measures for the further evaluation methods of article quality.
As an example for evaluating information quality the in- and out-degree values of an article can be taken. In the case of Wikipedia one article links to an other article only if it is somehow semantically related to it (Kamps and Koolen, 2009). However, Kamps and Koolen come to the conclusion that the link structure of Wikipedia follows a power-law distribution, similar to the link structure of the World Wide Web (Newman, 2004). In their analysis they point out, that an article with a low in degree value can also be relevant. They come to the conclusion that the link structure in the Wikipedia network may be not a good measure to show the importance of an article (Kamps and Koolen, 2009). Therefore this Thesis takes also betweenness and clustering coefficient values into account as they tells more about the position and the relevance of an article in the network.

Underlying Paper The paper "*Network analysis of user generated content quality in Wikipedia*" by Ingawale, Dutta, Roy and Seetharaman is the base for this master thesis. The purpose of this paper is to find out whether high or low quality user generated content produce different connectivity structures in Wikipedia. For this reason the authors use six different Wikipedias: Cebuano, Bosnian, Interlingua, Tagalog, Croatian and Slovenian. They took *featured articles* as a measure for high quality of an article. The main goal was to find out if these featured articles lie at strategic good positions in the network, so called *structural holes* (see section Structural Holes Model or Bridging Model). For their investigations they use social network analysis

(Wasserman and Faust, 1994), because they represented Wikipedia as a network of interactions between contributors (Ingawale et al., 2013). They introduced two different graph representations for presenting the Wikipedia graph. First, they made a user network, where each node is a contributor that has at least contributed once to an article and each edge between two articles means, that two contributors have contributed at least once to the same article. Second, they made the article network, where each node is an article and each edge between two nodes mean, that these articles share at least one common contributor. The experiments and results are presented in chapter Methodology.

3.3. Information Quality in Wikipedia by Content and Edit Statistics

It is also possible to define the quality of an article by looking only at the properties of an article itself, and ignoring its position in the network and who has contributed to the article. One way to do so is to look at the reputation of an article (Lih, 2004):

- ***Rigor:*** Counts the all edits made to an article. Articles that have undergone more revisions than others are thought of to be of better quality as it has been revised more often.
- ***Diversity:*** Counts the all unique users. Article quality can profit from more contributors as they have a high chance that they also provide distinct opinions on a topic.

The study of Lih concentrated on the quality of an article *before* and *after* it was cited in the press. He found out, that after an article was cited in the press more users contributed to this article and improved it. However, he also found out that there is often a small set of contributors that edit a specific article, which was also stated by Laniado and Tasso. The contribution count follows a power law distribution, where most contributions (24%) are done by only 6% of the contributors. This number even rises for featured articles where 21% of edits are done by only 2% of contributors (Stvilia et al., 2005). Also the chance that an article is edited by a different contributor during

one hour lies at 7% and raises up to 22% during 24 hours (Buriol et al., 2006).

Featured articles are 18 times larger in terms of median length than non featured articles (Stvilia et al., 2005). Therefore another possibility to define a good quality article is to measure the *word count* (Blumenstock, 2008). The author states that there are several advantages for using *word count*:

- the word count of an article is quite easy to get
- the calculation is easy and faster than other approaches like taking more measures into account (see Stvilia et al., 2005)

Blumenstock tested his approach by randomly taking 9513 random articles and 1554 featured articles.Then he declared every article with more than 2000 words as featured, and those with less than 2000 words he considered as non-featured articles. With this algorithm he was able to get a 96.31% accuracy.

class	n	TP rate	FP rate	Precision	Recall	F-measure
Featured	1554	0.936	0.023	0.871	0.936	0.902
Random	9513	0.977	0.064	0.989	0.977	0.983

Table 3.1.: Word count perfomance taken from Blumenstock, 2008

However, he also states that the word count only holds if featured articles are the measure for good quality articles. Otherwise he only showed that "*long articles are featured, and featured articles are long*" (Blumenstock, 2008). Blumenstock use only unbalanced data sets, which means that the number of featured and non-featured articles varies heavily. However, there is also research taking balanced data sets, where the number of featured and non-featured articles is the same (Lex et al., 2012). Lex et al. propose *factual density* as a metric to define the informativeness of an article. *Factual density* is the ratio between *fact count*, which is the number of facts within an article, and the size of an article. Lex et al. show that word count performs better with an unbalanced data set, which is biased towards non-featured article, but factual density performs better for balanced datasets. They were able to get an accuracy of 87.14% on their binary classification with a naive Bayes classifier on featured and non-featured articles.

3.4. Problems by Evaluating Information Quality

Due to the fact that anyone can edit (almost) every article, wikipedia is also vulnerable to vandalism. These vandalisms can be distinguished into five different types (Viégas, Wattenberg, and Dave, 2004):

- **Mass deletion:** the whole content of an article is deleted
- **Offensive copy:** vulgarities are added to the content
- **Phony copy:** text that is not relevant to the article is added
- **Phony redirection:** redirect pages are manipulated to lead to a different article
- **Idiosyncratic copy:** text is added to the article that somehow relates to the topic, but is not useful for the article

According to Viégas, Wattenberg and Dave the smallest revert times were for obscene edits with a mean time of 1.8 days and a median time of 1.7 minutes and the largest revert times were for complete deletions with a mean time of 22.3 days and a median time of 90.4 minutes. However, vandalism on featured articles are turned back faster with a mean time of 199 minutes and a median time of 9 minutes (Stvilia et al., 2005). These vandalisms are a problem of all information quality metrics introduced before. It can change the collaboration structure as the user now appears as a contributor to that article. It can change the network structure as for example a phony copy can change the in- and out-degree values as well as the position in the network. And it can also change the article structure as both the rigor and diversity values increase leading to a false better value of the quality of the article.

Another big problem to this thesis is the unbalanced number of featured and non-featured articles. Since labelling articles as featured is a community process, this takes some time. There are in general much more non-featured than featured articles, which will make the classifier biased towards non-featured articles.

4. Methodology

This chapter outlines the Methodology and the theoretical background which is the base of the experiments and the discussion of the results in chapter Experiments & Results. In order to understand and reproduce the experiments and results, the principles of Social Network Analysis are introduced as this is important to understand the four different networks that are generated from the Wikipedia dump. Afterwards Connectivity describes two different models - structural holes and closure - as these are used to find high quality articles in the article network and collaboration network. The different metrics that are used in this thesis are described in section Metrics. These metrics are then used to classify articles as featured and non-featured. The theoretical background on classification is discussed in section Classifier.
This thesis tries to find answers to the research questions stated in Research Questions. In order to find answers, the Wikipedia dump was downloaded and the data set was preprocessed to get rid of unnecessary data like redirects, disambiguations and edits that were made by bots. From this data set four different networks are generated on which different experiments were executed. Then the results were combined in order to get the best classifier. The following figure 4.1 summarizes the described steps.

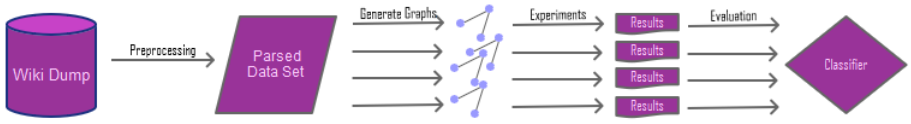


Figure 4.1.: Summary of steps necessary for this thesis

This thesis is based on the paper of Ingawale et al., who use six different language Wikipedias. They chose those different Wikipedias mainly because

of their small sizes and their relative young age. For answering their research question - whether featured articles lie at structural holes - they use the clustering coefficient and the average path length. They formulated the following hypothesis:

- "H_{CC} : *Featured Articles have lower clustering coefficients than non-Featured Articles.*"
- "H_{AP} : *Featured Articles have lower average path-lengths than non-Featured Articles.*"

As a result they get that for all six language Wikipedias both the clustering coefficient and the average path length are smaller for featured articles than for non featured articles. As an example of these results the team outlines the values for the featured article "Osama bin Laden" in the Cebuano Wikipedia. This article has a clustering coefficient of 0.58, where the average clustering coefficient of all articles is 0.88, and an average path length of 3.89, where the average path length of all articles is 4.13. This article therefore spans a structural hole connecting for example ""Fatawa", "Iran", "Yasser Arafat", "Mehiko", "Israel", even "Symphony No.5 (Beethoven)" and "Dragon Ball". The team concludes that articles that lie at structural holes can be expected to be of better quality (Ingawale et al., 2013).

This master thesis takes this paper as a base but uses the Norwegian Wikipedia instead. The aim is to reproduce this results and additionally try to find metrics with which articles can be automatically classified into featured and non-featured articles.

4.1. Social Network Analysis

As this thesis uses graph theoretical networks and metrics to automatically classify articles into featured and non-featured articles, the theory of Social Network Analysis is introduced here.
The field of Social Network Analysis evolved from the interest of affiliations between different individuals (Wasserman and Faust, 1994) and were developed by researchers of the field of social theory. Researchers wanted to gain information about social structures and its behaviour in "*political,*

economic, or social structure environment" and therefore a new model had to be introduced, which is called *Social Network Analysis (SNA)*. Since *SNA* focuses on the interactions between contributors, researchers tried to find new ways to describe these networks without relying on existing methods like from statistics (Wasserman and Faust, 1994). The idea of *Social Network Analysis* is to find methods which describe networks with its related notions (Wasserman and Faust, 1994).

There are a few important concepts in the theory of *Social Network Analysis* (Wasserman and Faust, 1994):

- Actors are linked together with relations.
- Activities between actors are dependent on each other, and not seen as individuals with no correlation between each other.
- Links between actors act as channel of information.
- A network of contributors can benefit or be constrained by the network structure.
- Links between actors can be of different structures and are seen as long lasting arrangements.

SNA benefits from the different perspectives of the different fields of study. The early developers of *SNA* found heavy use for mathematical models. The most important model for this thesis is graph theory. A graph representation of a network consists of nodes, which are connected by lines. Lines (in this thesis called edges) can either be *directed* or *undirected* which leads to a directed- or undirected graph. In the case of directed edges, one node is the source and another node is the target. In the case of undirected edges every node is source and target at the same time. Such a graphical representation is called a *sociogram*.

In the case of this master thesis there are four different networks generated and analysed. The two types of these networks are introduced now.

Social Network A *social network* merges nodes and edges, so that nodes are related to other nodes via edges. In a network there can be several sets of nodes, which can have different types and are related to each other (Wasserman and Faust, 1994). If there is only one set of node then this is called *one-mode network*. If there are two sets of nodes then this is called *two-mode network*.
During the initial stage of the Wikipedia dump analysis, the focus was set on the article network, which is created by articles and their links to other articles. The article network was generated in order to prove that high quality articles are more likely to connect different categories than normal articles. Additionally a contributor network was generated to show the contributions of users to featured and non-featured articles.

- **One Mode Network:** A one-mode network consists only of one set of nodes, for example *articles*. An edge between two nodes can be established if for example *one article links to another article*. The article network is generated for the English and Norwegian Wikipedia and, since articles link to other articles, is directed.

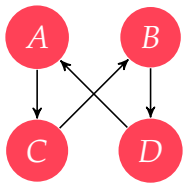


Figure 4.2.: Article Network

- **Two-Mode Network:** A two-mode network consists of two sets of nodes, for example *articles* and *contributors* (see figure 4.3). An edge between two nodes must be established between a node of the first set and a node of the second set in order "to be truly two-mode" (Wasserman and Faust, 1994).
  Such a relation can be for example *an article was edited by a contributor*. The two-mode network is undirected and is constructed for the Norwegian Wikipedia only.

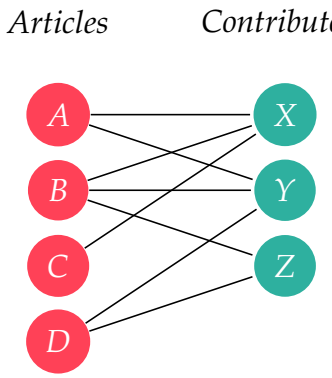


Figure 4.3.: Two-Mode Network of Articles and Contributors

- **Projection of Two-Mode Network:** Two-mode networks can be brought to a one mode network, where the nodes stay the same and the edges become the events. The resulting networks are called *projections*. In the case of this master thesis such a projection of the two-mode network was done to get the *Collaboration Network* and the *User Network*:
  - **Collaboration Network:** Two *articles* are connected if both of them were edited by the same *contributor*.
  - **User Network:** Two *users* are connected if both of them contributed to the same *article*.

  The projection of the two-mode network into the collaboration network was done in order to reproduce the results of Ingawale et al. However, the projection resulted in a drastic increase in the number of nodes and edges of the projected network, which we were not able to process with the given hardware resources. Therefore, the required metrics were calculated directly on the two-mode network instead.

4.1.1. Connectivity

This thesis tries to find high quality articles in Wikipedia by analysing its position in the different networks that are generated. High quality articles come in form of *featured articles* and the idea of Ingawale et al. is that these articles are *better connected* with other articles of the network. Therefore a

metric of what it means to be better connected has to be introduced. In theory two important models - *structural holes model* and *closure model* - were developed and are introduced now.

Structural Holes Model or Bridging Model A node is considered to be a *bridge* if removing it means that the shortest path length raises drastically. A node is called a local bridge of degree k if removing it the shortest path will be of length k. If removing a bridge means that two neighbour nodes are not able to meet any more then this is called a *Structural Hole* (Borgatti, 2010) or a *cutpoint* (Wasserman and Faust, 1994). This model implies a "think outside the box" mentality and suggests that an edge is more likely to be added between two nodes, if they do not share much other nodes yet (Gao, Hinds, and Zhao, 2013). From a social perspective, this means, that a worker does not necessarily have the same group of co-workers again and again but he can gain more information by being connected with many different groups (Burt, 2001; Gao, Hinds, and Zhao, 2013). In the case of Wikipedia this means, that an article node that lies at a structural hole in the article- or two-mode network does not only focus on its associated category, but also takes information from other categories into account. The idea of Ingawale et al. is that featured articles try to detail all aspects of a topic and therefore include more articles and from other categories as well. Therefore featured articles are thought to lie at structural holes. Ingawale et al. use the collaboration network as their network, together with the average path length and local clustering coefficient as their metrics. The reason why they use these metrics is, that featured articles are better connected and therefore have a lower average path length. The local clustering coefficient is also lower, because featured articles connect article from different categories. Therefore it is most likely that these articles are not connected leading to a lower clustering coefficient. This thesis uses the article- and two-mode network to find structural holes. For the article network the local clustering coefficient together with either the betweenness or average path length are used. For the two-mode network the metrics extended clustering coefficient and redundancy are used. The mentioned metrics are described in section Metrics.

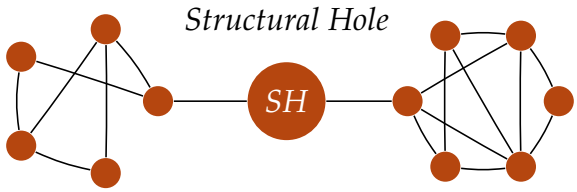


Figure 4.4.: Structural Hole

Closure Model or Bonding Model The closure model is contrary to structural holes model. This model suggests that nodes tend to connect with other nodes if they lie in the same cluster of the network (Gao, Hinds, and Zhao, 2013). If a network consists of many nodes that are connected within the cluster, then it means that information between nodes is shared more effectively, because there are more paths crossing these nodes than other nodes. It suggests also that an edge is more likely to be added if the two nodes share some other nodes yet (Gao, Hinds, and Zhao, 2013). A *closure model* also suggests a dense network, as information is only shared with nodes that are connected yet (Burt, 2001). According to the contrary nature of the closure model to the structural holes model, this model was not used to find high articles as the structural holes model provides the same information gain.

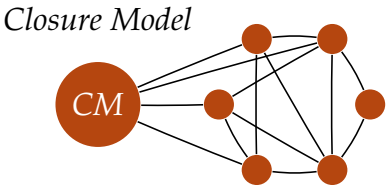


Figure 4.5.: Closure Model

4.2. Metrics

This section outlines the different metrics used to answer the research questions stated in chapter Introduction and Research Questions. At first, the density is described in order to compare the Norwegian and English

Wikipedia. Afterwards the degree centrality, betweenness, local/extended clustering coefficient, redundancy and average path length are used to compare featured and non-featured articles.

4.2.1. Density of a Graph

The density of a directed graph is calculated by dividing the number of all present edges by the number of all possible edges. The density Δ of a *directed graph* is

$$\Delta = \frac{L}{g(g-1)} \tag{4.1}$$

where L is the number of all present edges in the directed graph, and g is the number of present nodes in the directed graph. Δ is between 0, where no edges are present and 1, where every node is connected to every other node (Wasserman and Faust, 1994).

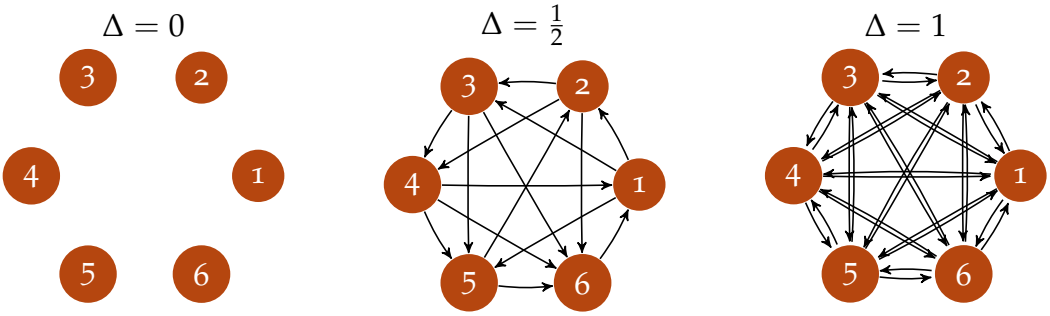


Figure 4.6.: Density of a Directed Graph

In the case of an *undirected graph* the density Δ is calculated as

$$\Delta = \frac{2L}{g(g-1)} \tag{4.2}$$

where L is the number of all present edges in the undirected graph, and g is the number of all present nodes in the undirected graph. Δ is again between 0, where no edges are present, and 1, where every node is connected with every other node.

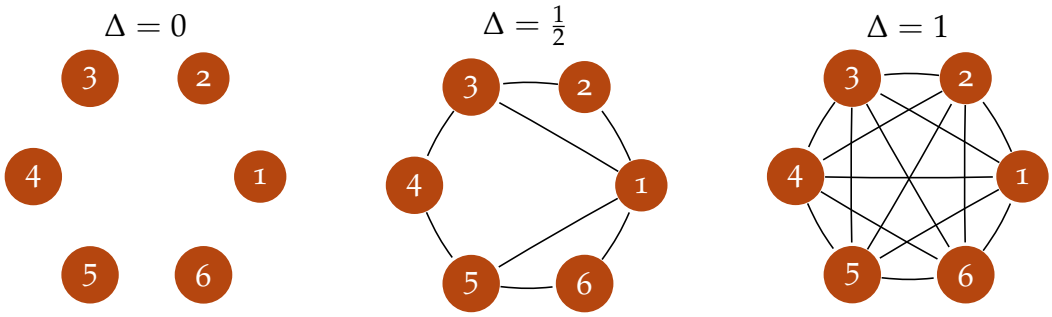


Figure 4.7.: Density of a Undirected Graph

4.2.2. Degree Centrality

Featured articles are thought to be better connected and reference more content from different topics than non-featured articles, they must have a higher degree centrality than non-featured articles. This metric is calculated for all four generated networks.

The *degree centrality* value measures how many other nodes in the network can be reached or how many other nodes in the network reach the node. In the case of undirected network the *in-degree centrality* C_{DI} is equal to the *out-degree centrality* C_{DO}, but this is not true for *directed graphs*:

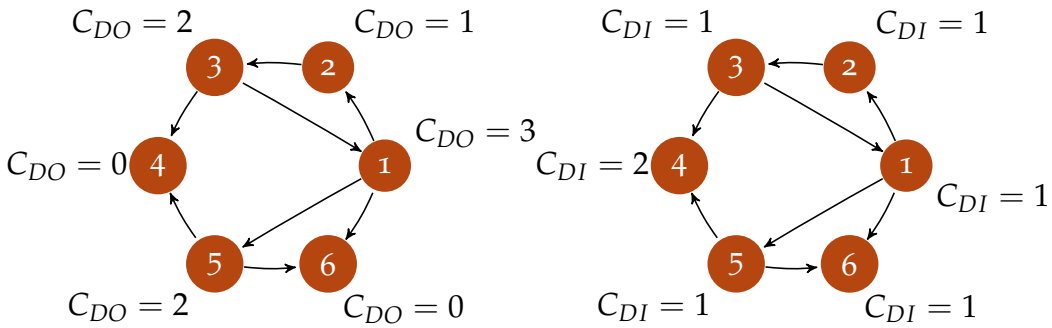
$$C_{DO}(n_i) = d_O \tag{4.3}$$
$$C_{DI}(n_i) = d_I \tag{4.4}$$

where d_O or d_I is the number of outgoing or incoming edges of node n_i (Wasserman and Faust, 1994).

The degree centrality for an *undirected graph* is calculated similarly:

$$C_{DO}(n_i) = C_{DI}(n_i) = C_D(n_i) \tag{4.5}$$

The degree values for the undirected graph are the sum of the in-degree and out-degree values for the directed graph.



Out-Degree In-Degree

Figure 4.8.: Directed Graph Degree

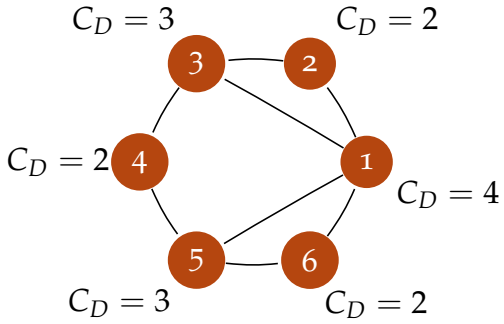


Figure 4.9.: Undirected Graph Degree

4.2.3. Average Path Length

The average path length is used together with the clustering coefficient to find structural holes in the network. It is also used by Ingawale et al. to find high quality articles as these articles are better connected with articles from different categories.

Nodes that are well connected with many other nodes have a lower *average path length*.

$$C_{APL}(n_i) = \frac{1}{|n| - 1} \sum_k d(n_i, n_k) \tag{4.6}$$

where $|n|$ is the number of all nodes in the network and $d(n_i, n_k)$ is the shortest distance from node i to node k.

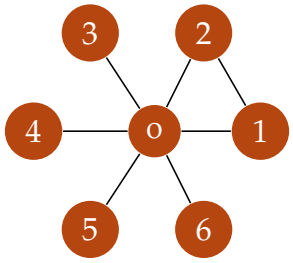


Figure 4.10.: Average Path Length

In the network of figure 4.10 node 0 has average path length of 1, which means that this node is connected with all other nodes, whereas node 3 has average path length of 1.8.

4.2.4. Betweenness Centrality

As stated by Ingawale et al. featured articles might lie at structural holes. A structural hole is indicated by a lower average path length and a lower clustering coefficient for high quality articles. As it was initially assumed that it was not possible to calculate the average path length for our big dataset out of the box by graph-tool, the betweenness centrality was chosen instead as it has a similar interpretation to average path length. Nodes that lie on many shortest paths have a high *betweenness centrality*. This means that nodes with a high betweenness centrality value lie relatively central in the graph and the chance is high, that a random path between two nodes will walk through this node.

$$C_B(n_i) = \sum_{j<k} \frac{g_{jk}(n_i)}{g_{jk}} \tag{4.7}$$

where g_{jk} is the number of all shortest paths from node j to node k, and $g_{jk}(n_i)$ is the number of all shortest paths from node j to node k that go though node i (Wasserman and Faust, 1994; Freeman, 1977).

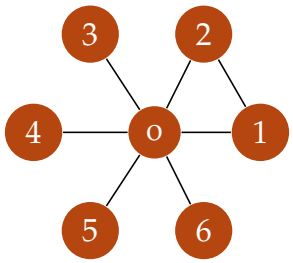


Figure 4.11.: Betweenness Centrality

In the network of figure 4.11 nodes 3 to 6 have betweenness centrality of 0, nodes 1 and 2 have a value of $\frac{1}{6}$ and node 0 has a value of $\frac{5}{6}$.

4.2.5. Local Clustering Coefficient

In order to prove the results of Ingawale et al. on featured articles lying at structural holes, the local clustering coefficient is also relevant. This metric is only used for the article and user network. The collaboration network was too big to be calculated with the available resources and the two-mode network needs a different metric as there are no direct links within articles or users.

The *local clustering coefficient* measures how well connected neighbours of a node are. If the clustering coefficient is high it is also an indication for a small-world behaviour (Barrat and Weigt, 2000):

$$C_{CC}(n_i) = \frac{1}{k_i(k_i-1)} \sum_{j \neq k, j,k \in N_i} e_{jk}(n_i) \tag{4.8}$$

where e_{jk} is an indicator function with value 1, if there is a link between j and k, and 0 if there is no link, and N_i are all neighbours of n_i.

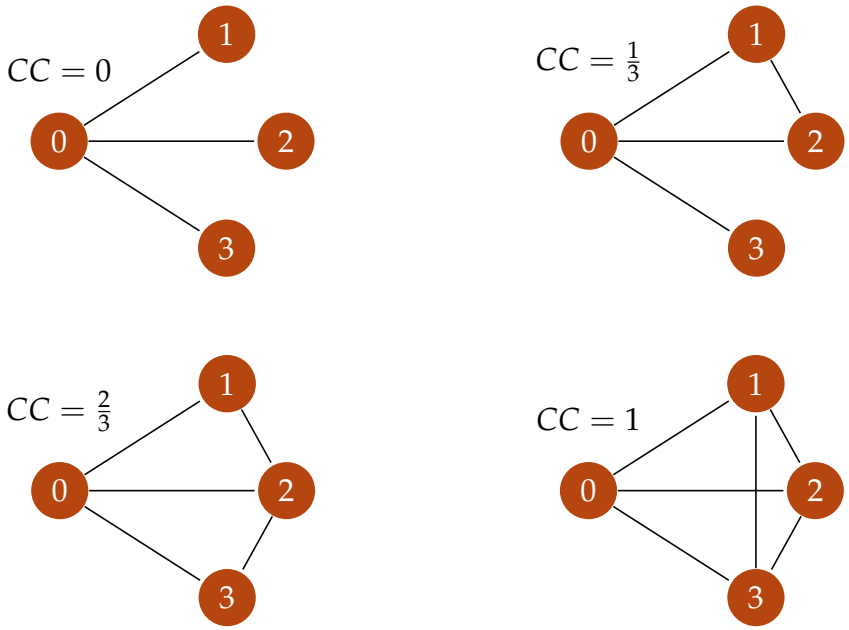


Figure 4.12.: Local Clustering Coefficient

4.2.6. Extended Clustering Coefficient

As mentioned in local clustering coefficient a different metric for two-mode network was required to measure the connectedness of neighbours of a node. For this reason the extension of the local clustering coefficient is introduced and is calculated in the experiments in order to find structural holes and to classify articles.

The *extended clustering coefficient* measures how well connected neighbours with a given distance of a node are. If the extended clustering coefficient is high then the neighbours with distance d are highly connected to each other. This measure is useful when calculating the clustering coefficient for two-mode networks as there is no edge within a set (Abdo and Moura, 2006; Xiao et al., 2007).

$$C^d(n_i) = \frac{|\{\{v,w\} : v,w \in N(n_i) | d_{G(V\setminus\{u\})}(v,w) = d\}|}{\binom{|N(n_i)|}{2}} \tag{4.9}$$

where $|N(n_i)|$ is the number of neighbours of node n_i, and $d_{G(V\setminus\{u\})}$ represents the set of vertices of neighbours n_i which have distance equal to d.

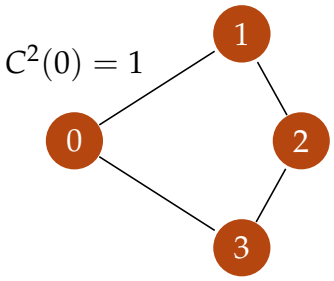


Figure 4.13.: Extended Clustering Coefficient

The extended clustering coefficient with distance 2 for node 0 is 1 because the two neighbours 1 and 3 have distance 2 (see figure 4.13).
For the two-mode network in this thesis a distance $d = 2$ was chosen, as there are two sets of nodes - article and user.

4.3. Classifier

After evaluating all the different metrics for all four networks the resulting values were used to classify articles as featured or non-featured articles. This is useful as featured articles are nominated by users and are then marked as featured or not. With a classifier, articles can be nominated as featured beforehand and users only have to agree or disagree on the suggestion. The task is to state for any article whether or not it is featured. Therefore a binary classification is used, since there are only two groups to which an article can belong - featured or not featured. A major problem is the unbalanced data set, with much more non-featured than featured articles. This leads to a biased result towards non-featured articles. This bias can be mitigated by balancing the data set by selecting an equal number of randomly chosen non-featured articles to featured articles. The different results between the balanced and unbalanced data set is shown in section Classification. For automatically classifying articles based on the calculated metrics the free software *WEKA* (Hall et al., 2009) is used.

The classifier tries to distinguish featured articles from non-featured articles. For any given new article it tries to find the class the article belongs to. For this reason the classifier uses 10-fold cross validation with naive Bayes model. This model assumes that metrics of all nodes in a network are independent from another. In order to calculate the contingency table and the metrics like recall and precision the machine learning tool *WEKA* is used (Hall et al., 2009).

4.3.1. Precision, Recall, F1 & ROC

In order to measure the correctness of a classifier the metrics *precision* and *recall* are used. Therefore the numbers of true and false positive, true and false negative classified items are written in a *contingency table*.

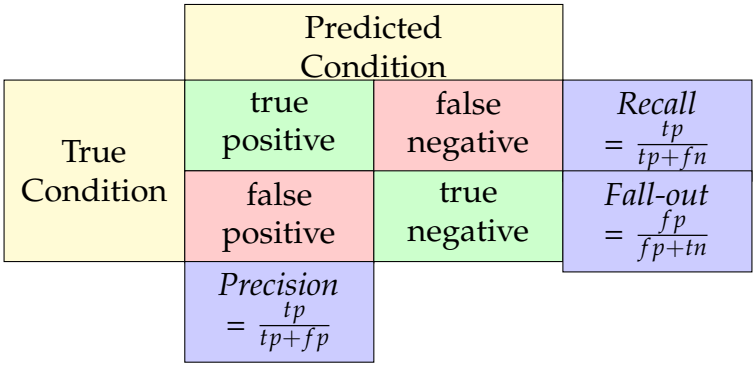


Figure 4.14.: Contingency Table

Featured article can either be *true positive*, if they were classified correctly, or *false negative*, if they were classified wrongly. Non-featured articles can either be *false positive*, if they were wrongly classified as featured, or *true negative*, if they were correctly classified as non-featured articles.

Recall or sometimes called *sensitivity*, is the fraction of relevant items that are returned dived by the number of items that should have been returned (Powers, 2011). Since there are many articles in Wikipedia but very little featured articles the bias is high for the unbalanced data set. For this reason

this thesis focuses on recall as this is more meaningful for this specific task than the precision.

Precision or sometimes called *confidence*, is the fraction of relevant items that are returned divided by the number of all items that are returned (Powers, 2011). Since there are many more non-featured articles, the precision is biased towards non-featured articles for the unbalanced data set. Due to completeness it will also be listed in the appropriate contingency tables.

F1-score is the harmonic mean of recall and precision and is calculated by $2 * \frac{recall * precision}{recall + precision}$. It therefore does not take the number of true negatives into account (Powers, 2011).

Fall-out is the fraction of non-relevant items that are returned divided by the number of all non-relevant items (Powers, 2011). This metric is important for the ROC area.

ROC area This analysis takes the fall-out rate and plots it against the recall rate. The resulting plot indicates whether the classifier was able to correctly classify the items or not. In the best case the curve is left skewed, indicating that the recall is high and the fall-out is low. The worst case is a heavily right skewed curve, meaning that the classifier returned many non-relevant items and little to none relevant items (Powers, 2011). Also, when the ROC value is high, then the curve is heavily left skewed, which is means a good classification result.

CCI stands for Correctly Classified Instances and is a metric on how well the classifier predicted the articles overall. It is calculated by dividing the sum of true positives and true negatives by the number of all articles, which is $\frac{tp+tn}{tp+fp+fn+tn}$. This metric is only meaningful for the balanced data set, as this set contains the same number of featured and non-featured articles.

5. Experiments & Results

5.1. Dataset Acquisition

Wikipedia is quite generous with the data it shares. It is very simple to download a *Wiki Dump* in a specific language and work with the data. It allows the user to get full database dumps of the last nine month in various file formats. For simplification reasons these formats are explained for the English Wikipedia (*WikiDumps*) but are the same for all different language Wikipedias. Wikipedia offers two main formats to process the Wiki data: *Extensible Markup language- XML* and *Structured Query Language -SQL* files. For the XML file format it also offers to download either a *7z*, *bz2* or *gz* file. For this thesis the *7z* compressing format was chosen. It compresses better than *bz2*, but does not provide a reliable error protection mechanism. In fact even with the higher possibility of damage in the downloaded files taken into account, the difference in file size and download speed was more then worth the risk. Nonetheless expecting damage or transfer errors in the downloaded dump files, it was decided to add a verification step to the data acquisition process. It turned out that the data dumps were never damaged.

5.1.1. XML Dumps

At the beginning of this thesis the idea was to work with the SQL dump of Wikipedia, because of its easy access to the data. For example SQL enables the user to perform flexible queries which makes it easy to export different parts of the dataset to be used in experiments. The data structures in the database can also be used to save preliminary results and to accumulate the

data from further analysis in a centralized way. The SQL-statement *Select* can then be used to get the filtered datasets. The SQL dump can be inserted to a MySQL database, but it will expand to about 320GB with current pages (all namespaces) and pagelinks within the English Wikipedia. Another problem is that the SQL JOIN operation, which simply aggregates data from different tables, takes a long time on such big datasets. Finally the restoration of the database from the dumps was very slow. Inserting the Wikipedia data into the database took about two weeks without any indexes and inserting the indexes took another two weeks. Initially only a standard PC hardware was available, and the execution of simple *Join* queries took too much time. After a few weeks of trying to insert indexes to speed up the queries, the decision to use the XML dump and parsing the needed data instead was made.

5.1.2. Preprocessing Wikipedia Dumps

As the data extraction process is always traversing the XML datafile in a linear fashion, random access to the data is not as important as fast data processing capabilities in general. Therefore XML data stream parsing provides faster results, especially on big data sets with the entire article revision history. The initial goal was to reduce the size of the dataset by removing all unnecessary data. Wikipedia uses the XML representation as an export format in order to provide a standardized form which can be easily read by humans. XML by itself contains repetitive markup and therefore needs much more disk space than other representations like the MySQL tables. But different to SQL it will not be expanded by indexes that have to be introduced in order to speed up the database access time. To further reduce the required disk space of the dataset, the XML files were processed to a *csv* file which were then loaded into the graph processing tool *graph-tool*. This process is described in chapter Construction of Data Files.

5.1.3. Data Set of Norwegian Wikipedia Dump

The following chapter outlines the size and the data processing process associated with the preparation of the datasets used by the experiments. Initially the dataset of the Norwegian Wikipedia is described in detail and the process of extracting the nodes and edges.

The Norwegian Wikipedia XML dump consists of a compressed file containing 107 GB of uncompressed XML. The file was downloaded and verified for correctness by checking the hash checksum.

Dump Verification Wikimedia enables the user to check the consistency by providing SHA1 and MD5 checksums for all dump files that can be downloaded. In order to ensure a correct transfer, the user can use the given hash algorithm to verify that the received file is identical to the one on the server, by checking if the given hash code matches the hash code produced by a rerun of the hash calculations. The Cygwin environment tools for the verification of the checksums were used to verify the consistency of the files (*Cygwin*). As none of the dump files showed any errors, it can be safely assumed that the transfer was completed correctly.

The following section describes the XML schema and the strategy which was used to preprocess the dataset.

```
<?xml version="1.0" encoding="utf-8" ?>
<mediawiki>
  <page>
    <title>Testpage</title>
    <ns>0</ns>
    <id>1</id>
    <revision>
      <id></id> <!--Revision ID != Page ID-->
      <parentid></parentid> <!--OPTIONAL-->
      <timestamp></timestamp>
      <contributor>
        <username>JohnDoe</username>
        <id></id>
      </contributor>
      <comment></comment>
```

```
      <text xml:space="preserve"></text>
      <sha1></sha1>
      <model></model>
      <format></format>
    </revision>
    <revision>
      <id>2</id>
      <parentid>1</parentid>
      ....
    </revision>
  </page>
</mediawiki>
```

The Wikipedia dumps essentially consist of a list of page elements. This page elements represent the articles of the Wikipedia. Each page element can be categorized into different namespaces by the "ns" element, it has a page id, which is a unique integer number, and a title containing the article name. The content of the article itself can be found in the revisions attached to each page. The last revision is the current state of the article. Each edit of the Wikipedia leads to a new revision being created.
The revision tag contains the text of the article in the *text* element. Additionally it provides information about the contributor who initiated the edit, the timestamp of the edit and the parent revision id which is the base for the current revision. Each revision is also identified by its own id field which is unique for all revisions of the current page.

As XML is a structured markup language the content of the individual tags can only be understood regarding the context of the tag. For example the ID tag in the page XML element (/Page/id) does have a different meaning than the id tag of the revision element for example (Page/revision/id), even if it has the same XML tag name. In order to cope with this context specific meaning of the tags in XML, the parser has to be aware of the current context and has to interpret current tags value according to the current context. As the size of the XML file is rather large, it is necessary to be able to parse the file without loading the whole XML file into memory. Therefore a stream parser, which processes the file tag by tag, and keeps track of the current context is required.

Speeding up the parser Additionally to the parsing of the XML file, the SQL dump of the Wikipedia is used. The advantage of this second data-source is that it provides a list of article names and ids during the parsing of the XML dump. This enables the parser to verify and look up all ids of link target articles at a time where the parser hasn't already processed the target article itself. This allows the parser to complete the entire processing in one pass.
In order to obtain this list of articles and their ids, the SQL dump of the page table (page.sql.gz) was downloaded. The result is the page table containing the names of the articles and their ids. Then only the articles which are in namespace 0and are no redirect pages where selected and exported into a csv file. Which in turn was then fed into the parser to serve as a article name to id lookup table. As this table consists only of names and id, it can be held in memory for fast access.

5.1.4. Construction of Data Files

During the parsing of the XML stream, the text of the current revisions of all articles is analysed and all local Wikipedia links are detected. Each link is checked against the pages.csv file created from the MySQL dump. If an article with a name matching the name in the link can be found, the link is pointing to a valid Wikipedia article. Finally a new edge can be created consisting of the id of the article as the sourceid and id of the page the link pointed to from the pages.csv file. The new found edge is then saved to the edges.csv file. As the linktext in Wikipedia is case sensitive, there can be two article names only differing in the casing of for example the first letter, pointing to two completely different articles.

Processing Data Set As discussed in the previous chapters, the parser has to be context aware. Furthermore it has to be fast and cannot be allowed to keep too much data in memory. Especially it must not load the entire data file in memory because this would quickly exhaust the memory and cannot be done practicably.

To cope with this requirements a state machine design was used for the parser. The state machine has flag(state) variables for the page and revision context. The simplified algorithm can be described by the following steps:

1. **Load Symbol:** Load the first XML Symbol from the streamparser
2. **Abort if at the end:** If the XML symbol is *End Document* abort immediately.
3. **Is Start Element?** If the event type is *Start Element* note the type in the current context variable.
   a) **Is Page Element:** Parser is now entering page context
   b) **Is Revision Element** Parser is now entering revision context
4. **Is End Element:**
   a) **Is Page Element:** process the page data, save it to the nodes file, append the edges file end clear the page cache
   b) **Is Revision Element** An entire revision has been parsed. Process the Links in the content and save all useful data to the files. Clear the revision cache (see section 5.1.4).
   c) **Other Element:** If the Element contains useful data save the current content variable using the name of the current XML element as the key. Clear the current content variable
5. **Characters:** Append the new characters to the current content string variable
6. **Load Next Symbol** Get the next XML symbol from the streamparser and continue with Step 2

The state based parsing of the algorithm described in the steps above allow for elements which have the same name but different content or meanings depending on the context they are encountered. This enables the parser to process a structured hierarchical XML markup in a linear fashion (see 5.1 on page 45).

Processing Revision Content When a complete revision has been parsed more steps are required. The following paragraphs outline the content analysis as performed by the parser when a revision is being processed.

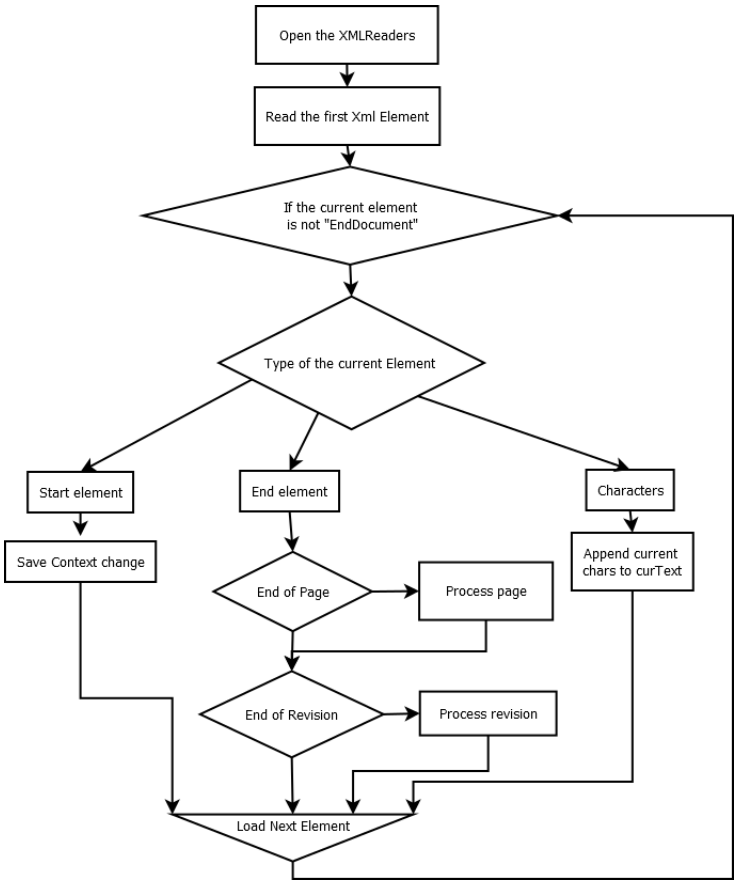


Figure 5.1.: Flow Chart of Data Preprocessing

Detecting Disambiguations If a disambiguation page is detected the entire page can be omitted. Disambiguation are marked by a {{disambiguation}} or {{dab}} tag in the Wikipedia mark-up of the page. If such a tag is detected the parsing of the content of the revision is aborted and the page is omitted.

Detecting Redirects If the content of the revision indicates that this page is a redirect to another Wikipedia article then the entire page can also be omitted.

Detecting outgoing Wikipedia links When it is ensured that the page is really an article the parsing of the content of the article can be initiated. The goal is to detect all outgoing links which point to other valid and existing Wikipedia articles.
As Wikilinks have the mark-up `[[Linktarget|Link Target Name]]` all links need to be found and the *Linktarget* string needs to be extracted from the link mark-up in order to find the corresponding id in the lookup table.

This can be done as indicated in the following steps:

1. **Search:** Find the first occurrence of [[
2. **Search:** Find the first occurrence of]] after first position of [[
3. **Extract:** The link target name
4. **Process:** Find corresponding id in lookup table
5. **Remove:** all text from 0 to the position of]] from the content of the article
6. **If not empty Continue:** If there is still content in the article continue with step 1

Filtering Data Set In order to reduce the file size of the dataset several filters where defined. The parser uses special tags which are not included in the resulting dataset, as criteria to further reduce the size of the acquired data. The implemented stream parser uses the following rules to filter the XML stream:

- Namespace 0 Filter: Only pages having a namespace value of 0 are relevant, as only namespace 0 pages are articles.
- Revision Timestamp Filter: Only revisions newer than 11.01.2014 will be included in the dataset.
- Size Filter: Only revisions which have more than 512KB of texts
- Type Filter: All redirect or disambiguations pages will be ignored

5.1.5. Structure and Size of Generated Data Files

During the processing of the XML dump the parser generates three output files. The *node file*, the *edge file* and the *users file*. This chapter outlines the file

format and the size of the 3 output files of the Norwegian Wikipedia. Further it will point out the reduction in size relative to the original Wikipedia dump. As all the files contain tabular data, a simple file format containing plain text values separated by delimiters was chosen.

Node File The node files contains the articles and their revision meta information. Each record consists of a single line.

Position	Name	Delimiter	Description
1	node ID	|	Number identifying the unique ID of the page
2	is featured	;	"true" if the article is featured, "false" if not
3	contributor ID	;	Number identifying a unique contributor by its unique ID
4	text length	;	Number indicating the length of the article text in characters
5	time stamp	\	String representation of the time stamp of the revision
> 5	repeat 2-5	\	For each revision a set of column is appended to the line

Table 5.1.: Node File Format

As a result the node file is easy to parse and load into other graph processing tools and looks like follows:

```
nodeid|isfeatured;contributorid;textlength;timestamp\...
nodeid|isfeatured;contributoid;textlength;timestamp\...
```

To retrieve data from this node file format one has to go through the following steps for each line:

1. Split at | to get the node ID and the rest of the line
2. Split at \to get all the revisions from the rest of the line
3. Split at ; to get the individual fields of the revisions

Edge File The Edge file represents all the links between the articles. It only contains the links of the most current revision of each article.

Position	Name	Delimiter	Description
1	source ID	|	Number identifying the unique ID of the source page
2	target ID	newline	Number identifying the unique ID of the target page

Table 5.2.: Edge File Format

The edge file can be easily split by the | delimiter to get the individual fields

```
sourceid\targetid
```

Usernames File The user file contains the mapping of all extracted contributor names and their ids.

Position	Name	Delimiter	Description
1	user ID	|	Number identifying the contributor by its unique id
2	user name	newline	String identifying the contributors name

Table 5.3.: Username File Format

```
userid\username
```

Statistic of Norwegian Wikipedia Due to the relevance for this thesis the statistic for the Norwegian Wikipedia is presented here. Since only a few experiments are made on the English Wikipedia the corresponding statistics is presented were needed.
For the experiments the data dump of the Norwegian Wikipedia from

November 2nd, 2015 was chosen. The dumps consists of all pages and its revision that were made from the beginning of the Norwegian Wikipedia. Only pages which had edits in the time interval from 2014-11-01 until 2015-11-02 were chosen and consisted of the following:

	all	largest component
Articles	162,853	127,644
Edges	3,075,788	2,684,306
Featured Articles	229	226
Featured Articles Edges	1,869	1,851

Table 5.4.: Article Network Statistics of Norwegian Wikipedia

Featured articles edges are edges that connect featured articles and no non-featured article.

	all (no bots)	bots
Contributors	9,857	12
Revisions	479,234	493,339

Table 5.5.: User Network Statistic of Norwegian Wikipedia

For this Wikipedia dump twelve bots were identified and all its revisions were not added to the parsed data set. The bots are responsible for more than 50% of all revisions (which is the same as edits), as there are *BjornNbot* (5277), *Chobot* (142), *DanmicholoBot* (1525), *EmausBot* (2531), *HaakonBot* (9), *JAnDbot* (3), *JhsBot* (130), *KjelloBot* (0, no edit from 2014-11-01 on), *LA2-bot* (1), *PladaskBot* (0, no edit from 2014-11-01 on), *SDBot* (73), *RussBot* (2), *Jeblad (bot)* (483646).

5.2. Experiments with Article Network

The *article network* consists of articles as nodes which are linked together if there is at least one reference in the text content of the article to another article in namespace 0.
For example the text content of article A:

```
This article is very interesting and has much to tell.
But this article does not include every piece of information
therefore this article links to articles B,D and F where
further information on the specific topic is provided.
```

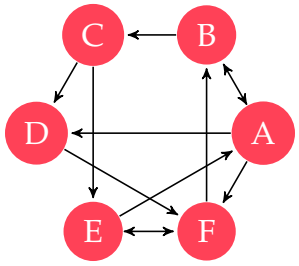


Figure 5.2.: Article Network Illustration

Figure 5.2 is an illustration on how the article network is constructed. Given the example text above node *A* links to nodes *B*, *D* and *F*. This graph also shows that nodes *B* and *E* also link to article *A*.
The article network is the only network where the experiments are done for both Norwegian and English Wikipedia. For the other networks only the Norwegian data set is used.

In order to answer research questions 1 & 2 - whether featured articles have different properties than non-featured articles and whether they lie at structural holes - the in-, out- and total degree distribution is calculated and also the betweenness, clustering coefficient, average path length and average text length are determined. At first the statistics of the article network for the Norwegian and English Wikipedia is presented. Afterwards the previously mentioned metrics are calculated and interpreted.

5.2.1. Statistics

The article networks for the Norwegian and English Wikipedia have following statistics:

	Norwegian	English
Articles	162,853	4,553,106
Edges	3,075,788	99,251,199
Featured Articles	229	3,146
Largest Component		
Articles	127,644	3,639,273
Edges	2,684,306	85,292,893
Featured Articles	226	2,996
Density	0.0164%	0.00064%
Pseudo Diameter	17	72

Table 5.6.: Article Network Norwegian and English Wikipedia Statistics

The largest component for both the Norwegian and English Wikipedia contains most featured articles, but nevertheless there are only 0.17% in the Norwegian and 0.08% in the English Wikipedia articles labelled as featured (see table 5.7). Also the density is quite small for both Wikipedias and the pseudo diameter, which is the shortest longest paths between two nodes, is smaller for the Norwegian than for the English Wikipedia. The reason might be because of the bigger size and the age of the English Wikipedia. At some point there are so many articles, that the users simply do not link to each and every article in the network any more leading to a longer diameter.
In the Norwegian Wikipedia there is one featured article that has a much higher in degree than the other featured articles. This article happens to be the article about *Oslo*, which is the capital of Norway.

5.2.2. Degree Distribution

The following section outlines the in-, out- and total degree distribution of the article network for the Norwegian and English Wikipedia. It is expected that the distributions will follow a power-law, which means that there are a lot of articles with low degree and few articles with a very high degree. The plots are given on a log/log scale. If the curve is similar to a straight line, then the distribution follows a power-law. The standard deviation is not very useful in this case as the power-law benefits outliers. However, due to completeness the standard deviation is also given.

	Norwegian	English
In Degree		
All Articles	$\mu = 21.03, \sigma = 0.49$	$\mu = 23.44, \sigma = 0.19$
Featured Article	$\mu = 366.45, \sigma = 1363.05$	$\mu = 214.85, \sigma = 78.82$
Non-Featured Article	$\mu = 20.41, \sigma = 162.43$	$\mu = 23.28, \sigma = 374.32$
Out Degree		
All Articles	$\mu = 21.03, \sigma = 0.14$	$\mu = 23.44, \sigma = 0.02$
Featured Article	$\mu = 174.22, \sigma = 119.71$	$\mu = 126.59, \sigma = 78.82$
Non-Featured Article	$\mu = 20.76, \sigma = 48.82$	$\mu = 23.35, \sigma = 44.67$
Total Degree		
All Articles	$\mu = 42.06, \sigma = 0.54$	$\mu = 46.87, \sigma = 0.2$
Featured Article	$\mu = 540.67, \sigma = 1441.92$	$\mu = 341.44, \sigma = 844.03$
Non-Featured Article	$\mu = 42.17, \sigma = 181.82$	$\mu = 46.63, \sigma = 384.02$

Table 5.7.: Article Network Norwegian and English Wikipedia Degree Distribution

Table 5.7 above shows the detailed values for the degree distribution, whereas figure 5.3 below provides a graphical representation of the mean and standard deviation values.

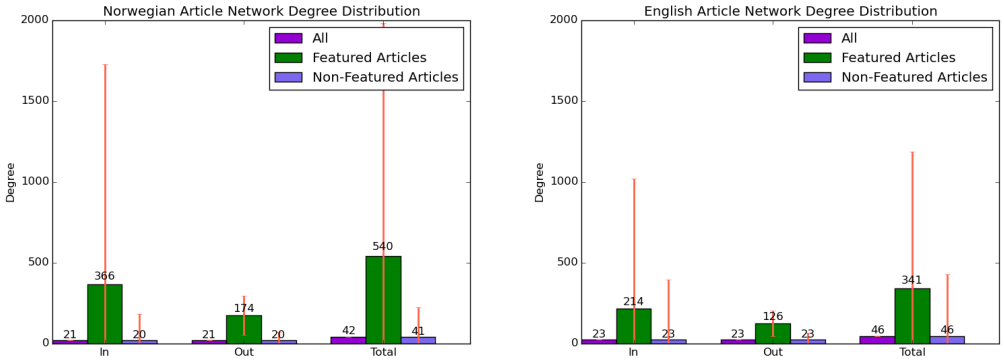


Figure 5.3.: Article Network Degree Distribution

Total Degree Distribution The total degree distribution for both Norwegian and English Wikipedia article network are plotted on a log/log scale and show a power-law distribution. This means that articles with a high total degree distribution are more likely to increase their degree further than article with a low total degree distribution.

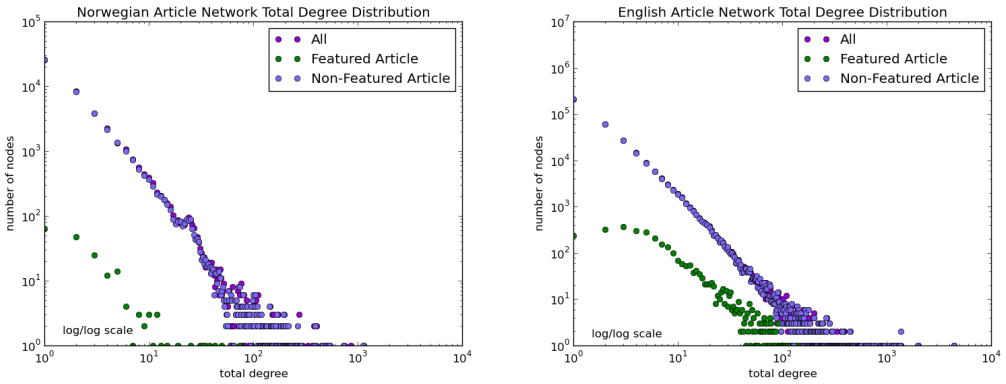


Figure 5.4.: Total Degree Distribution; Left: Norwegian Wikipedia, Right: English Wikipedia

In Degree Distribution The in degree distribution for the article network for the English and Norwegian Wikipedia are plotted on a log/log scale and show a power-law distribution. It seems that featured articles have a lower in degree than non-featured articles, but since there are less featured articles with a low in degree, the average in degree is higher for featured than for non-featured articles.

Out Degree Distribution The out degree distribution for the article network for the English and Norwegian Wikipedia are plotted on a log/log scale and show a power-law distribution for non-featured articles. Featured articles seem not to have a power-law distribution. It seems that featured articles in general have a high out degree, and that there are few featured articles with few outgoing links and many articles with many outgoing links.

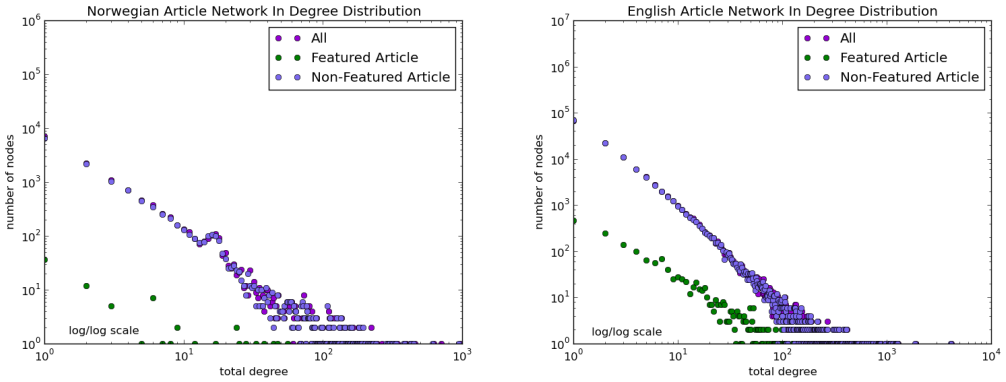


Figure 5.5.: In Degree Distribution; Left: Norwegian Wikipedia, Right: English Wikipedia

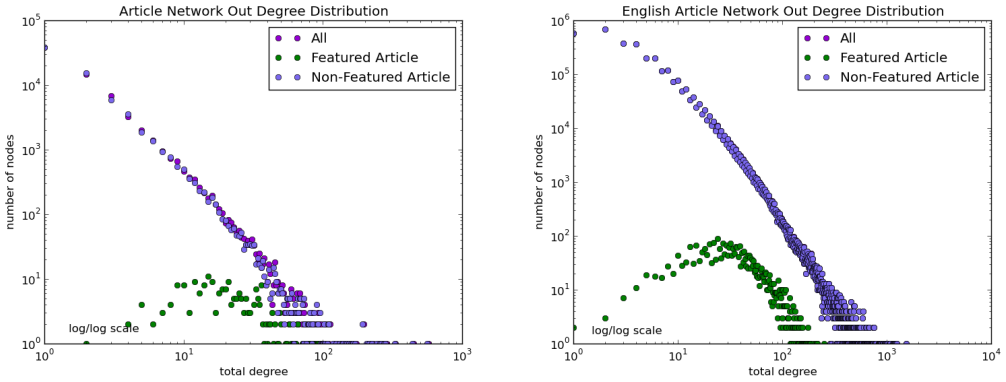


Figure 5.6.: Out Degree Distribution; Left: Norwegian Wikipedia, Right: English Wikipedia

The degree distribution partially answers the first research question - whether featured articles have different properties than non-featured articles. This can be confirmed for in-, out-, and total-degree distribution on their average values. But for the out-degree distribution also the form of the curve is different, also confirming the research question.
It is interesting that the form of the curves in any case for the Norwegian Wikipedia is very similar to the curves for the English Wikipedia. Even though the English Wikipedia is older and much bigger the properties seem to scale with the size of the Wikipedia.

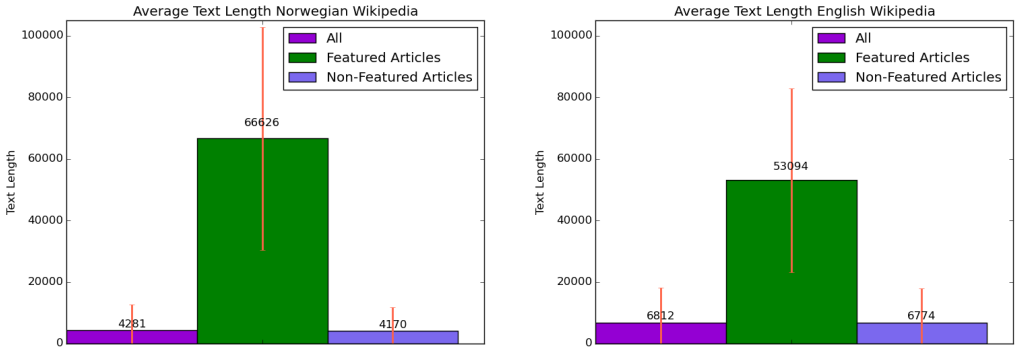


Figure 5.7.: Article Network Average Text Length; Left: Norwegian Wikipedia, Right: English Wikipedia

5.2.3. Average Text Length

One way to define quality in Wikipedia is to measure article length (see section 3.3). Since featured articles contain a lot of information, they might be longer in terms of text length than non-featured articles (Blumenstock, 2008). The correlation on text length and article quality is examined here. The text length is measured in characters per article.

	Norwegian	English
All Articles	$\mu = 4281, \sigma = 8297$	$\mu = 6812, \sigma = 11348$
Featured Articles	$\mu = 66626, \sigma = 36207$	$\mu = 53094, \sigma = 29884$
Non-Featured Articles	$\mu = 4170, \sigma = 7729$	$\mu = 6774, \sigma = 11242$

Table 5.8.: Average Text Length

As one can see the text length for featured articles is higher than for non-featured articles, but there is also a much higher standard deviation.

As discussed by Blumenstock mentioned in section 3.3 article length might be an indicator for good quality article. In the case of the Norwegian

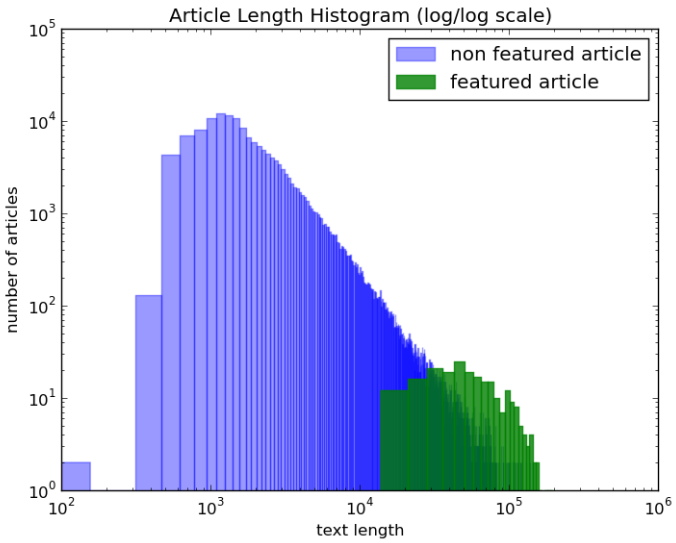


Figure 5.8.: Article Length Histogram for Norwegian Wikipedia

Wikipedia the average article length of featured articles is 17 times higher than for non-featured articles, whereas for the English Wikipedia the article length is eight times higher for labelled high quality articles. But as Blumenstock also mentioned it might only be the fact that featured articles are longer and not necessarily that long articles are of good quality. Average text length will be used for classification later.

5.2.4. Average Path Length

The average path length can only be calculated for the Norwegian Wikipedia as the runtime of this algorithm would take too long for the English Wikipedia (about 220 days).

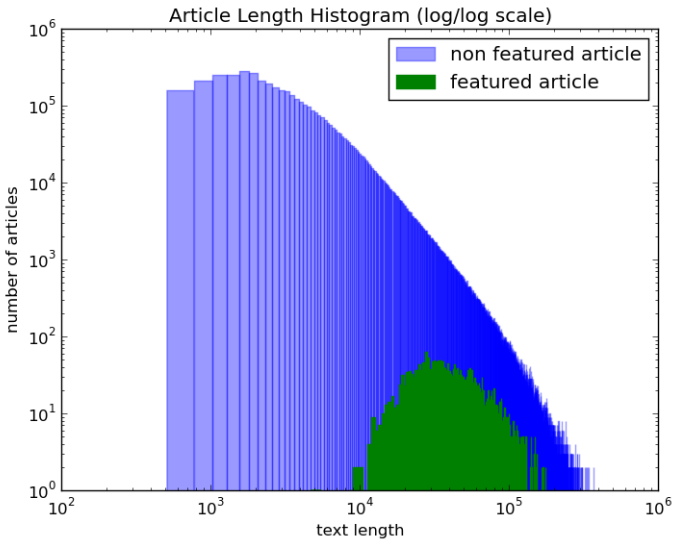


Figure 5.9.: Article Length for English Wikipedia

	Norwegian
All Articles	$\mu = 4.09, \sigma = 0.32$
Featured Articles	$\mu = 3.74, \sigma = 0.43$
Non-Featured Articles	$\mu = 4.09, \sigma = 0.32$

Table 5.9.: Article Network Average Path Length mean and standard deviation

Figure 5.10 shows the mean and standard deviation of the average path length for featured and non-featured articles.
The average path length for featured articles is shorter than for non-featured articles in the article network. This is part of the indication for a structural hole and partly answers the first two research questions - whether featured articles have different properties than non-featured articles and whether they lie at structural holes.

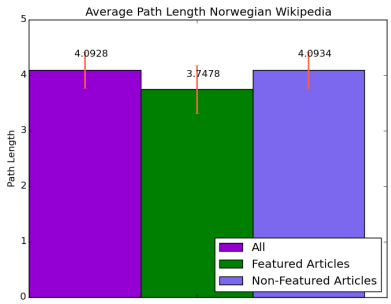


Figure 5.10.: Average Path Length for Norwegian Article Network

5.3. Experiments with User Network

The *user network* for the Norwegian Wikipedia consists of user nodes which are linked together if they collaborated on the same article. There can also be multiple edges from one user to the other if they collaborated on multiple articles together. If two users collaborated on a featured article, then the edge is considered *featured*. The resulting network is undirected and covers a timespan of one year of edits.
The user network does not answer any stated research questions, but it is interesting whether there are users that contribute more to featured articles than to non-featured articles. This means to find a correlation on the betweenness of a user and the number of articles the user collaborated with another user. This means that user that lie relatively central in the network are more likely to edit more on featured articles than other user.
At first the user network construction is explained, then the statistics and afterwards the correlation of featured article edge count and betweenness is introduced.
An example user network:

```
user 1: article A, article B, article C
user 2: article B, article D
user 3: article C, article E, article F
user 4: article A, article B
user 5: article E, article F
```

```
user 6: article G
```

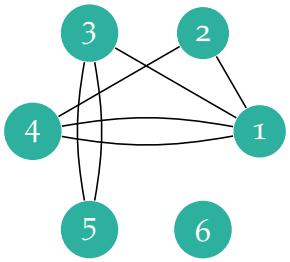


Figure 5.11.: User Network Illustration

5.3.1. Statistics

The user network for the Norwegian Wikipedia has the following statistics, and consists of contributions within one year:

	Norwegian
User	9,857
Edges	208,017
Featured Edges	5,691
Largest Component	
User	8,551
Edges	207,977
Featured Edges	5,691
Total Degree	$\mu = 48.64, \sigma = 5.11$
Density	0.42%
Pseudo Diameter	5

Table 5.10.: User Network Norwegian Wikipedia Statistics

In the case of user network, a *featured edge* means that two user contributed on a featured article together. There are more featured edges than featured articles because similar users contributed to different featured articles, which

leads to the high number of featured edges.
The user network is small compared to the other networks like article-, collaboration or two-mode network. Since the network considers only a timespan of one year of contributions there are 9857 registered and non-bot users. On average a registered user contributed to 49 different articles within a year.

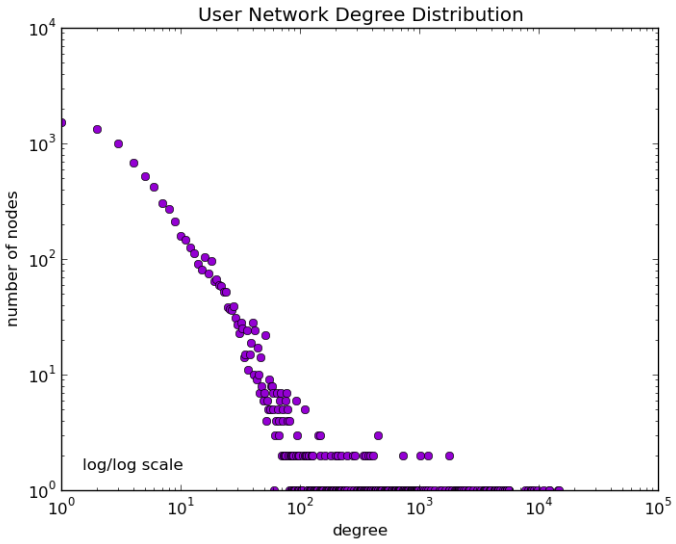


Figure 5.12.: User Network Degree Distribution

The degree distribution of the user network is plotted on a log/log scale (see figure 5.12) and follows a power-law distribution. This means that user that already contributed to many articles are more likely to contribute to even more articles. Whereas users that did not edit many articles yet are less likely to edit many articles.
For the experiment afterwards only the featured article edge count is needed, which is the number of edits to a featured article with another user.

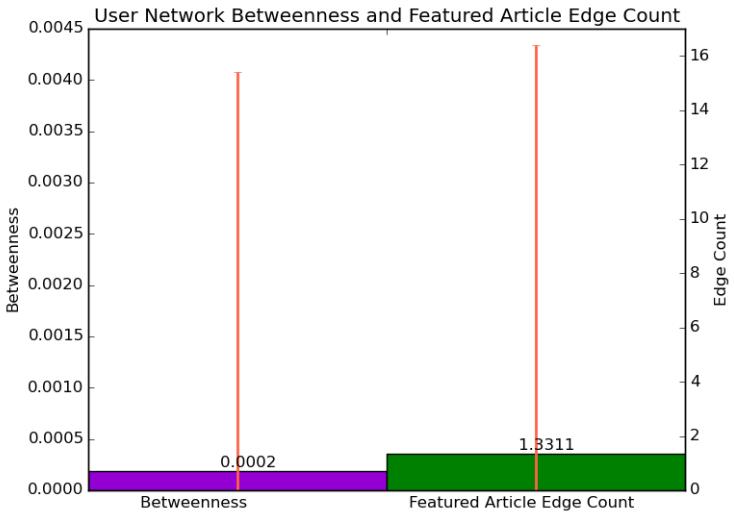


Figure 5.13.: User Network: Betweenness and Featured Article Edge Count

5.3.2. Featured Article Count

In the case of the user network two nodes are connected if both contributed to the same article. If this article is featured then the edge is considered a *featured edge*. Therefore every user can have a featured edge. Users that contributed to many featured articles together with similar other users then they have more featured edges. This number of featured edges belonging to a user is called *featured edge count*. Users with a high featured edge count therefore contributed to many featured articles, but do they also lie relatively central in the network? With other words, do users with a high featured edge count do have a high betweenness value? This question will be answered by looking at the largest connected component only.

As shown in figure 5.13 the average betweenness for a user is very small with 0.002 and a standard deviation of 0.0039 and the average featured article edge count has an average of 1.33 with a standard deviation of 15.08. These high standard deviations are caused to the power-law distribution

of both metrics. Since there are a few very high , but mostly rather small values, the standard deviation is very high.

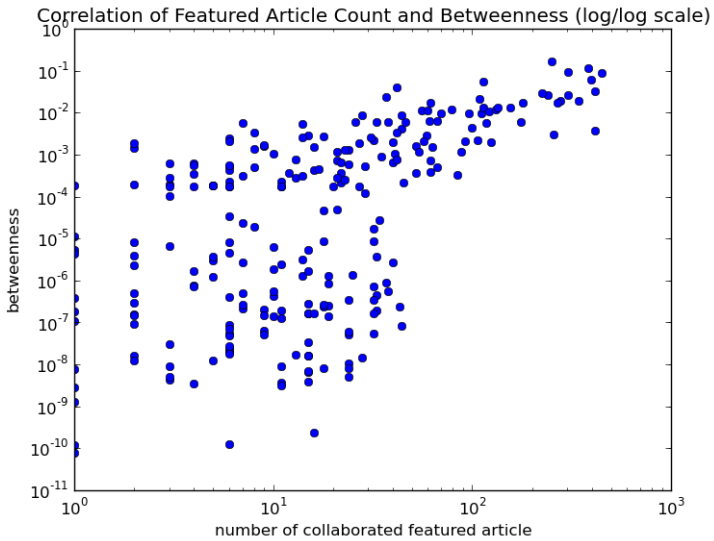


Figure 5.14.: Correlation Featured Edge Count and Betweenness for User Network

Figure 5.14 shows on a log/log scale that when the number of featured edge count increases the betweenness also increases. This means that users that contribute to many featured articles lie relatively central in the network. The Pearson correlation is 0.70 whereas the Spearman correlation has a value of 0.29. Pearson correlation measures linear dependencies, whereas Spearman measures monotonic relationships. The major problem here is that Pearson correlation needs normal distributed values, but in the case of the user network the degree- and betweenness values are power-law distributed and therefore the Spearman correlation should be taken. Since the Spearman correlation has only a value of 0.29 the correlation between featured article edge count and betweenness is very weak. So users that lie relatively central in the network do not necessarily contribute to more featured articles than other users. This value might also be caused to the short timespan of contributions.

5.4. Experiments with Collaboration Network

Since Ingawale et al. focuses on collaboration networks in their paper, this network is also presented here. However, due to its size it was not possible to calculate all relevant metrics in the given time. The collaboration network is therefore only presented due to completeness reasons and this master thesis does not focus on it any further.
The *collaboration network* consists of article nodes that are linked together if two articles were edited by the same user. Due to runtime limitations all parallel edges have been removed, so that there is only one edge from article A to article B even if they have been edited by multiple same users. The resulting network is undirected.
An example for collaboration network:

```
article A: user 1, user 2, user 3
article B: user 2, user 4
article C: user 3, user 5, user 6
article D: user 1, user 2
article E: user 5, user 6
article F: user 7
```

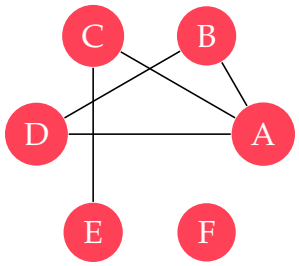


Figure 5.15.: Collaboration Network Illustration

5.4.1. Statistics

The collaboration network for the Norwegian Wikipedia has the following statistics:

	Norwegian
Articles	162,853
Edges	535,001,167
Featured Articles	229
Largest Component	
Articles	161,428
Edges	535,000,956
Featured Articles	229
Total Degree	
All Articles	$\mu = 6628, \sigma = 13$
Featured Articles	$\mu = 13055, \sigma = 11676$
Non-Featured Articles	$\mu = 6619, \sigma = 5457$
Density	4.1%
Pseudo Diameter	4

Table 5.11.: Collaboration Network Norwegian Wikipedia Statistics

Since this network is quite dense compared to the other networks it was not possible to perform any of the algorithms necessary to evaluate the properties of this network. The algorithms used for this thesis mostly have a quite long runtime and a high memory consumption. Therefore the algorithms are run on the Two-Mode network which constructed this collaboration network as a one mode projection.
There are even some advantages when looking on the original two mode network, for example the projections of two mode networks are much denser than its original, since every node of the two-mode network produces $\frac{d(d-1)}{2}$ edges in the projection (Latapy, Magnien, and Vecchio, 2008). Since there are many edges induced due to the projection, the clustering coefficient is higher and might not be significant (Latapy, Magnien, and Vecchio, 2008). Therefore *redundancy* for two-mode networks was introduced by Latapy, Magnien, and Vecchio. This metrics will be used in the next section 5.5.

5.5. Experiments with Two-Mode Network

As described in Methodology the two-mode network is used instead of the collaboration network, which is used by Ingawale et al. to find featured articles at structural holes.
The *two-mode network* consists of two different types of nodes. There are article nodes which are connected to user nodes only. There is no link between articles or between users. The edges represent edits made by users to articles within one year. The network contains only edits made by registered users.
An example for Two-Mode Network:

```
article A: user 1, user 2, user 3
article B: user 2, user 4
article C: user 3, user 5, user 6
article D: user 1, user 2
article E: user 5, user 6
article F: user 4
```

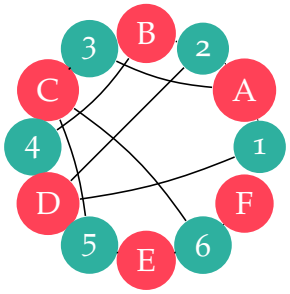


Figure 5.16.: Two-Mode Network Illustration

5.5.1. Statistics

The two-mode network for the Norwegian Wikipedia has the following statistics:

	Norwegian
All Nodes	172,709
Articles	162,853
Users	9,856
Edges	479,234
Largest Component	
All Nodes	169,979
Articles	161,428
User	8,551
Edges	477,116
Density	0.03%
Pseudo Diameter	12

Table 5.12.: Two-Mode Network Norwegian Wikipedia Statistics

	Norwegian
All Articles	$\mu = 2.96, \sigma = 8.76$
Featured Articles	$\mu = 18.83, \sigma = 48.50$
Non-Featured Articles	$\mu = 2.93, \sigma = 8.56$
User	$\mu = 55.8, \sigma = 601.46$

Table 5.13.: Two-Mode Network Total Degree Distribution

The number of all nodes in the two-mode network is the sum of all articles and all users that contributed within a year. Also the number of edges is the number of contributions within a year.

5.5.2. Degree Distribution

The degree distribution for the two-mode network is presented here as it shows the contribution behaviour of user and also the number of edits made to articles within a year.
As table 5.13 shows, the total degree is relatively small for articles. Non-featured articles have on average 3 contributions, whereas featured articles have about 19 contributions within a year. However, the standard deviation

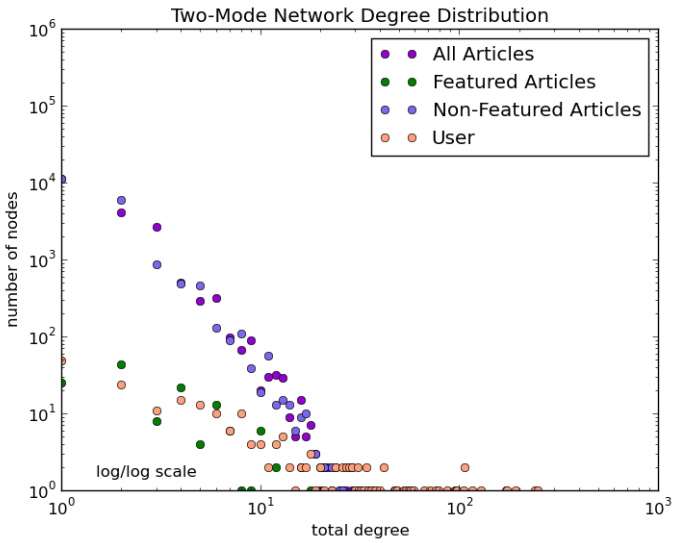


Figure 5.17.: Two-Mode Network Degree Distribution

for featured articles is 49, which means that there are only a few featured articles that received many contributions leading to a high average. As figure 5.17 suggests, there are many articles with less and a few articles with many contributions.
A user made on average 56 contributions to articles, but again, the standard deviation is very high with 601 meaning that there are a few that made a lot of contributions, whereas the majority did much less edits.
These results are characteristics for a power-law distribution and therefore also an indication for a small-world property where every node is connected by only a small number of steps.
The degree distribution partially answers the first research question - whether featured articles have different properties than non-featured articles. This question can be confirmed for degree distribution where featured articles show a different average and standard deviation than non-featured articles.

5.5.3. Average Redundancy

The local clustering coefficient is only meaningful in non bipartite graphs, since the direct neighbours can be connected. In bipartite graphs neighbours of a node can not be directly connected. Redundancy takes this into account and measures the connectedness of neighbours with a distance of 2.

Norwegian	Redundancy
All Nodes	$\mu = 46.81, \sigma = 172.63$
Articles	
All Articles	$\mu = 49.25, \sigma = 176.81$
Featured Articles	$\mu = 50.58, \sigma = 116.28$
Non-Featured Articles	$\mu = 49.25, \sigma = 176.88$
User	$\mu = 0.81, \sigma = 0.7$

Table 5.14.: Two Mode Network Redundancy

The values for articles are much higher than those for users. This is because there are about sixteen times more articles in the network than users. In the case of articles alone there is no big difference whether or not an article is featured. The mean values and the corresponding standard deviations are quite similar. The values for each node are used for classification and the results and usefulness of redundancy are discussed there (see section 5.6.2).

5.5.4. Average Path Length

The average path length should be less for featured articles than for non-featured articles as this is an indication for a structural hole.
Table 5.15 shows that the average path length for featured articles is indeed less for featured articles than for non-featured articles.
Figure 5.18 shows that there are many non-featured articles with a low average path length, but also that there many non-featured articles with a high average path length. This is the reason why on average the path length for featured articles is lower than for non-featured articles.

Norwegian	Path Length
All Nodes	$\mu = 3.75, \sigma = 0.54$
Articles	
All Articles	$\mu = 3.72, \sigma = 0.52$
Featured Articles	$\mu = 3.35, \sigma = 0.50$
Non-Featured Articles	$\mu = 3.73, \sigma = 0.52$
User	$\mu = 4.24, \sigma = 0.52$

Table 5.15.: Two Mode Network Average Path Length

This partially answers the first two research questions - whether featured articles have different properties than non-featured articles and whether featured articles lie at structural holes - and therefore also the fourth research question - whether the tow-mode network have similar differences in the properties for featured and non-featured articles. Ingawale et al. use clustering coefficient and average path length in order to confirm their hypothesis for their collaboration network. This thesis uses the two-mode network instead, but shows that featured articles do have on average a lower average path length than non-featured articles.

5.6. Classification

In order to answer the third and fourth research question - whether featured articles can automatically be classified - a classifier is trained. This classifier uses a naive Bayes model, as the assumption is that the used metrics are independent from each other. A 10-fold cross validation is used to train and test the data set, as the number of featured articles is low and with this validation method all articles are used as training and testing set. The classification uses a binary classifier, where the positive class is represented by featured articles and the negative class is represented by non-featured articles respectively. For the classifier the metrics betweenness, clustering coefficient and article length are used for the article network and for the two-mode network the metrics betweenness, redundancy and average path length are used. These properties were chosen as they were also used to answer the first two research questions. Only the article and two-mode

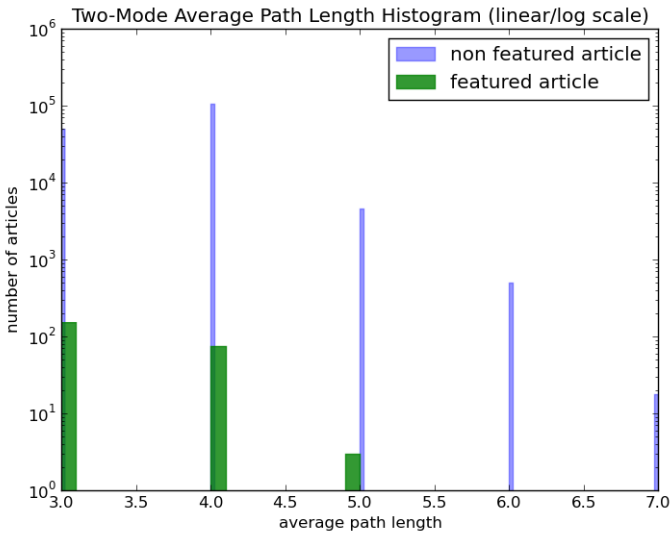


Figure 5.18.: Two Mode Network Average Path Length Histogram

network are used as the user network does not have featured articles nodes and the collaboration network is too big to be analysed in the given time.Two different data sets were chosen. First, the whole data set was taken, leading to biased results as there are many more non-featured articles than featured articles. Then the data set was created by randomly selecting the same number of non-featured articles as featured articles and the same tests as for the first data set were performed. The different tests and results are explained in the corresponding subsection.

5.6.1. Classification for Article Network

The following table lists the article length, local clustering coefficient and betweenness for featured and non-featured articles for the Norwegian and English Wikipedia.
This table shows that the local clustering coefficient is smaller for featured articles than for non-featured articles and that the betweenness is higher for featured articles than for non-featured articles. These results can be compared with those from Ingawale et al., even though the values are not

	Norwegian	English
Article Length		
Featured Articles	$\mu = 66677, \sigma = 36342$	$\mu = 52743, \sigma = 29866$
Non-Featured Articles	$\mu = 3819, \sigma = 7324$	$\mu = 6419, \sigma = 11021$
Clustering Coefficient		
Featured Articles	$\mu = 0.0006, \sigma = 0.0006$	$\mu = 0.0931, \sigma = 0.0625$
Non-Featured Articles	$\mu = 0.0039, \sigma = 0.0171$	$\mu = 0.3104, \sigma = 0.2769$
Betweenness		
Featured Articles	$\mu = 0.0148, \sigma = 0.0711$	$\mu = 1.3e^{-5}, \sigma = 6.2e^{-5}$
Non-Featured Articles	$\mu = 0.0003, \sigma = 0.0038$	$\mu = 6.8e^{-7}, \sigma = 3.4e^{-5}$

Table 5.16.: Article Network: Article Length, Clustering Coefficient, Betweenness

as high as for this paper (see chapter Methodology). This might be an indication for a structural hole in the article network, but since the standard deviation is higher than the mean value further investigation should be done to answer the second research question, whether featured articles lie at structural holes. Ingawale et al. use the complete edit history in their constructed collaboration network. The experiment here focuses on the article network of the current revision, which is a different network on a different timespan. In order to improve the article network to better fit the network Ingawale et al. used, the whole timespan of edits might be used. However, this will lead to a very big network for the Norwegian and English Wikipedia.
As explained in the Introduction to classification, in order to classify an article featured or non-featured a 10-fold cross validation with naive Bayes was performed. For this binary classification the positive class is represented by featured articles, and the negative class is represented by non-featured articles. The experiments for the article network are performed on the Norwegian and English Wikipedia for the article network and on the Norwegian Wikipedia only for the two-mode network. All experiments are performed with the complete data set (including all featured articles and all non-featured articles) and with a balanced data set (including all featured articles and the same number of non-featured articles as featured articles, which are randomly chosen).

Classification with Article Length Since article length is easy to calculate in the network it is the first to have a look at.
At first the contingency table is presented:

	Unbalanced				Balanced			
	Norwegian		English		Norwegian		English	
	FA	NFA	FA	NFA	FA	NFA	FA	NFA
FA	184	45	1321	1825	158	71	2016	1130
NFA	1634	160454	48732	4501228	20	209	1029	2117

Table 5.17.: Article Network: Contingency Table on Article Length

For the Norwegian Wikipedia unbalanced data set 184 out of 229 featured articles were classified correctly and 1634 were classified as featured even though they are not (see table 5.17). For the English Wikipedia unbalanced data set 1321 articles were classified correctly (out of 3146). For the Norwegian Wikipedia data set 158 featured articles were classified correctly, which is 14% less as for the unbalanced data set, but for the English Wikipedia balanced data set it classified 2016 featured articles correctly.
Now it is important to get recall, ROC values and the CCI, in order to quantify the quality of this classifier (see section Classifier):

	Unbalanced				Balanced			
	Norwegian		English		Norwegian		English	
	FA	NFA	FA	NFA	FA	NFA	FA	NFA
TP Rate	0.803	0.99	0.42	0.989	0.69	0.913	0.641	0.673
FP Rate	0.01	0.197	0.011	0.58	0.087	0.31	0.327	0.359
Precision	0.101	1	0.026	1	0.888	0.746	0.662	0.652
Recall	0.803	0.99	0.42	0.989	0.69	0.913	0.641	0.673
F1	0.179	0.995	0.049	0.994	0.777	0.821	0.651	0.662
ROC Area	0.994	0.994	0.98	0.98	0.925	0.925	0.738	0.738
CCI	98.9656%		98.8896%		80.1310%		65.6866%	

Table 5.18.: Article Network: Correctness of Classifier on Article Length

The precision is small for the unbalanced data set, because there are far more

non-featured articles than featured articles in the article network for the Norwegian and English Wikipedia. This is why the recall is more important in this unbalanced case as it only focuses on the classified featured articles and not on the classified non-featured articles (see table 5.18). However, for the balanced data set the recall is worse as for the unbalanced data set for the Norwegian Wikipedia. For the English Wikipedia the recall value is nearly the same for featured and non-featured articles. The precision in the balanced data set is higher for featured articles than for non-featured articles, as it is possible that some articles are considered as featured and not tagged as such yet. The percentage of correctly classified instances is high for the unbalanced data sets, which is not surprising and not meaningful, as there are many non-featured articles. Nevertheless, the CCI value for the balanced data set for the Norwegian Wikipedia is good, whereas for the English Wikipedia it is lower. As described by Blumenstock article length might be an indicator for featured articles, but not all long articles need to be featured.

Classification with Local Clustering Coefficient The local clustering coefficient is a metric on how well connected neighbours of a node are. The idea is that featured articles lie at structural holes and therefore should have lower clustering coefficient together with a lower average path length. On average featured articles do have lower clustering coefficients (see table 5.16) which might be used in order to classify articles.

	Unbalanced				Balanced			
	Norwegian		English		Norwegian		English	
	FA	NFA	FA	NFA	FA	NFA	FA	NFA
FA	0	229	0	3146	214	15	2420	726
NFA	0	162088	0	4549960	175	54	1496	1650

Table 5.19.: Article Network: Contingency Table on Local Clustering Coefficient

The classifier with only local clustering coefficient as metric did not classify any featured article correctly for the unbalanced data sets (see table 5.19). This might be, because the values are very small for both featured and non-featured articles. Surprisingly, it did classify most featured articles correctly

in the balanced data set for both the Norwegian and English Wikipedia. However, it also classified most non-featured articles as featured which leads to a low precision for featured articles (see table 5.20).

	Unbalanced				Balanced			
	Norwegian		English		Norwegian		English	
	FA	NFA	FA	NFA	FA	NFA	FA	NFA
TP Rate	0	1	0	1	0.934	0.236	0.769	0.524
FP Rate	0	1	0	1	0.764	0.066	0.476	0.231
Precision	0	0.999	0	0.99	0.550	0.783	0.618	0.694
Recall	0	1	0	1	0.934	0.236	0.769	0.524
F1	0	0.999	0	0.995	0.692	0.363	0.685	0.597
ROC Area	0.859	0.859	0.819	0.819	0.587	0.587	0.680	0.680
CCI	99.8589%		98.0960%		55.5153%		64.6853%	

Table 5.20.: Article Network: Correctness of Classifier on Local Clustering Coefficient

For the unbalanced data set of the Norwegian and English Wikipedia local clustering coefficient alone is not enough to classify articles. For the balanced data set most featured articles were classified correctly, but also many non-featured articles were also classified as featured, leading to a low CCI value of 55.52% for the Norwegian and 64.6853% for the English Wikipedia.

Classification with Article Length and Local Clustering Coefficient After classifying articles with their article length and local clustering coefficient alone, the next step is to combine these two metrics as article length might benefit from clustering coefficient.
For the unbalanced data set taking both metrics, article length and clustering coefficient results in 201 correctly classified featured articles (17 more than with article length alone - see table 5.17), but also in 583 wrongly classified non-featured articles for the Norwegian Wikipedia. For the balanced data set there are less wrongly classified non-featured articles, but also less correctly classified featured articles for both the Norwegian and English Wikipedia.

The classifier is biased towards non-featured articles for both the unbalanced

	Unbalanced				Balanced			
	Norwegian		English		Norwegian		English	
	FA	NFA	FA	NFA	FA	NFA	FA	NFA
FA	201	28	1485	1661	170	59	2638	508
NFA	2217	159871	56390	4493570	23	206	978	2168

Table 5.21.: Article Network: Contingency Table on Article Length and Local Clustering Coefficient

	Unbalanced				Balanced			
	Norwegian		English		Norwegian		English	
	FA	NFA	FA	NFA	FA	NFA	FA	NFA
TP Rate	0.878	0.986	0.472	0.988	0.742	0.900	0.839	0.689
FP Rate	0.014	0.122	0.012	0.528	0.100	0.258	0.311	0.161
Precision	0.083	1	0.026	1	0.881	0.777	0.730	0.810
Recall	0.878	0.986	0.472	0.988	0.742	0.900	0.839	0.689
F1	0.151	0.993	0.049	0.993	0.805	0.834	0.781	0.745
ROC Area	0.987	0.986	0.965	0.965	0.915	0.915	0.815	0.815
CCI	98.6169%		98.7250%		82.0961%		76.3827%	

Table 5.22.: Article Network: Correctness of Classifier on Article Length and Local Clustering Coefficient

and balanced data sets. The ROC area is better for the unbalanced data set, as the number of correctly classified non-featured articles is much higher as the number of correctly classified featured articles leading to a CCI of 82.09% for the Norwegian and 76.38% for the English Wikipedia.

Classification with Article Length and Betweenness As stated by Ingawale et al. the average path length for featured articles should be smaller than for non-featured articles (see chapter 4). In the case of this network betweenness takes the place for average path length due to time complexity run time reasons. In this case the betweenness should be higher for featured articles than for non-featured articles, since more shortest paths go through featured articles. For the Norwegian Wikipedia the betweenness is higher for featured articles than for non-featured articles (see table 5.16).
The combination of article length and betweenness classified 10 featured

	Unbalanced				Balanced			
	Norwegian		English		Norwegian		English	
	FA	NFA	FA	NFA	FA	NFA	FA	NFA
FA	174	55	1249	1897	127	102	3086	60
NFA	1500	160588	46683	4503277	18	211	2908	238

Table 5.23.: Article Network: Confusion Matrix on Article Length and Betweenness

articles less correctly than article length alone for the unbalanced data set of the Norwegian Wikipedia. For the Norwegian Wikipedia the balanced data set classified only 127 featured articles correctly, whereas for the English Wikipedia the balanced data set classified 3086 featured articles correctly. This bias in the balanced data set in the English Wikipedia for featured articles is interesting, as it can not be seen in the Norwegian Wikipedia. This might be due to the big size of the English Wikipedia as the betweenness might become very small for all articles in general.

	Unbalanced				Balanced			
	Norwegian		English		Norwegian		English	
	FA	NFA	FA	NFA	FA	NFA	FA	NFA
TP Rate	0.76	0.991	0.397	0.99	0.555	0.921	0.981	0.076
FP Rate	0.009	0.24	0.01	0.603	0.079	0.445	0.924	0.019
Precision	0.104	1	0.026	1	0.876	0.674	0.515	0.799
Recall	0.76	0.991	0.397	0.99	0.555	0.921	0.981	0.076
F1	0.183	0.995	0.049	0.995	0.679	0.778	0.675	0.139
ROC Area	0.993	0.993	0.98	0.98	0.894	0.894	0.797	0.797
CCI	99.0420%		98.9330%		73.7991%		52.8290%	

Table 5.24.: Article Network: Correctness of on Article Length and Betweenness

For the unbalanced data sets article length and betweenness is not a good combination for the classifier as it classified only 75% of featured articles in the Norwegian Wikipedia and 39% of featured articles in the English Wikipedia correctly. The balanced data set behaved the opposite way, where the English Wikipedia performed with 98% of correctly classified featured articles better than the Norwegian Wikipedia with only 55%. This behaviour was unexpected and might be discussed in some future work. However, the CCI shows the same behaviour for the balanced data set as for the

unbalanced data set. With a value of 73.80% for the Norwegian and 52.83% for the English Wikipedia the CCI values are rather small.

Classification with Article Length, Local Clustering Coefficient and Betweenness This classifier takes all three metrics - article length, local clustering coefficient and betweenness - in order to classify articles.

	Unbalanced				Balanced			
	Norwegian		English		Norwegian		English	
	FA	NFA	FA	NFA	FA	NFA	FA	NFA
FA	187	42	1404	1742	135	94	3023	123
NFA	1884	160204	53398	4496562	20	209	2744	402

Table 5.25.: Article Network: Contingency Table on Article Length, Local Clustering Coefficient, Betweenness

For the unbalanced data sets this classifier is between the classifier with article length alone and the one with article length and clustering coefficient as combination. It performs better for featured articles than the one with article length alone, and it performs better for non-featured articles than the one with article length and clustering coefficient as combination. For the balanced data set the number of correctly classified featured articles is less as for the unbalanced dataset for the Norwegian Wikipedia. For the English Wikipedia the number of correctly classified featured articles is about two times higher than for the unbalanced data set. This might again be caused due to the betweenness. Again some future work might analyse this difference for the Norwegian and English Wikipedia in betweenness.

Since there is a difference for the Norwegian and English Wikipedia in betweenness also the percentage of correctly classified featured articles differs for both Wikipedias. When not comparing the Norwegian and English Wikipedia and looking on the precision and recall itself, then the classifier worked well in classifiying featured articles for the unbalanced data set of the Norwegian Wikipedia and the balanced data set of the English Wikipedia. However, the CCI value for the balanced data set is better for

	Unbalanced				Balanced			
	Norwegian		English		Norwegian		English	
	FA	NFA	FA	NFA	FA	NFA	FA	NFA
TP Rate	0.817	0.988	0.446	0.988	0.590	0.913	0.961	0.128
FP Rate	0.012	0.183	0.012	0.554	0.087	0.410	0.872	0.039
Precision	0.09	1	0.026	1	0.871	0.690	0.524	0.766
Recall	0.817	0.988	0.446	0.988	0.590	0.913	0.961	0.128
F1	0.162	0.994	0.049	0.994	0.703	0.786	0.678	0.219
ROC Area	0.986	0.986	0.965	0.965	0.888	0.888	0.830	0.830
CCI	98.8134%		98.7890%		75.1092%		54.4342%	

Table 5.26.: Article Network: Correctness of Classifier on Article Length, Local Clustering Coefficient, Betweenness

the Norwegian Wikipedia with 75.11% than for the English Wikipedia with 54.43%.

Summary of Classification To sum up the best classifier to choose is the combination of article length and local clustering coefficient. It performs best with the recall and ROC area and betweenness is not needed. The third research question can now partially be answered. For the Norwegian Wikipedia it is possible to classify articles as featured or non-featured articles. The best approach is to take article length and clustering coefficient in a 10-fold cross validation with naive Bayes. The results for the English Wikipedia did not show these properties. This might be because there are many more non-featured articles than featured articles, or because many good articles are not tagged as featured yet, leading to a wrong interpretation of the results.

5.6.2. Classification for Two-Mode Network

As for the article network a naive Bayes classification is used. In this binary classification the positive class is represented by featured articles and the negative class is represented by non-featured articles. The experiments are evaluated by a 10-fold cross validation.

The classifications are made only for articles and not for users. Therefore only the nodes that are articles are considered in the classification process. The goal is to find a combination of metrics - average path length, betweenness and redundancy - that work best for the Norwegian Wikipedia to automatically classify articles as featured or not. The English Wikipedia is not considered here as the two-mode network was not generated.

The combinations

- average path length alone
- redundancy alone
- average path length and redundancy

did not classify any featured article correctly.
Therefore only the remaining combinations to classify articles automatically are shown now.

Classification with Betweenness The betweenness of node is a metric on how many shortest paths go through a node in comparison to all shortest paths.
This classifier (see table 5.27) table classified only 42 featured articles

	Unbalanced		Balanced	
	FA	NFA	FA	NFA
Featured Article	42	187	30	199
Non-Featured Article	1038	160161	22	207

Table 5.27.: Two Mode Network: Contingency Table on Betweenness

correctly and 187 wrong for the unbalanced data set. For the balanced data set it only had 30 featured articles correctly.
The results are better for the unbalanced data set (see table 5.28), as the number of non-featured articles is so much higher. Also the number of correctly classified articles is lower in the balanced data set. The classifier with only betweenness is biased towards non-featured articles and is not very good with a CCI of 51.75% for the balanced data set.

	Unbalanced		Balanced	
	FA	NFA	FA	NFA
TP Rate	0.183	0.994	0.131	0.904
FP Rate	0.006	0.817	0.096	0.869
Precision	0.039	0.999	0.577	0.510
Recall	0.183	0.994	0.131	0.904
F1	0.064	0.996	0.214	0.652
ROC Area	0.692	0.692	0.513	0.513
CCI	99.2411%		51.7467%	

Table 5.28.: Two Mode Network: Correctness of Classifier on Betweenness

Classification with Betweenness and Path Length The path length of a node is the average path length to any other node in the network in the largest connected component. Since betweenness and path length are both related to shortest paths for a node, the results in table 5.29 are quite similar to table 5.27 on the classification with betweenness alone.

	Unbalanced		Balanced	
	FA	NFA	FA	NFA
Featured Article	42	187	31	198
Non-Featured Article	1105	160094	25	204

Table 5.29.: Two Mode Network: Contingency Table on Betweenness and Path Length

The number of correctly classified featured articles is the same as for betweenness alone for the unbalanced data set and classified only one featured article more correctly. The classifier did suggest slightly more non-featured articles as featured as for the classification with betweenness alone. This is not surprising as the classifier with path length alone did not classify any featured article correctly.
For both the unbalanced and balanced data set the classification results on betweenness and average path length are not very good (see table 5.30). This might be due to the bias towards non-featured articles, for which the recall suggests good results. However, overall the results show that the classifier with betweenness and average path length for both the unbalanced and balanced data set does not classify featured articles automatically sufficiently.

	Unbalanced		Balanced	
	FA	NFA	FA	NFA
TP Rate	0.183	0.993	0.135	0.891
FP Rate	0.007	0.817	0.109	0.865
Precision	0.037	0.999	0.554	0.507
Recall	0.183	0.993	0.135	0.891
F1	0.062	0.996	0.217	0.646
ROC Area	0.729	0.729	0.552	0.552
CCI	99.1996%		51.3100%	

Table 5.30.: Two Mode Network: Correctness of Classifier on Betweenness and Path Length

Classification on Betweenness and Redundancy Redundancy shows how well connected a network is after removing a node. Therefore featured articles should have a higher redundancy than non-featured articles (see table 5.14). But since the mean values for featured and non-featured articles are quite similar and the standard deviation is high, the classification results are similar to the classification results with betweenness alone and betweenness combined with path length (see table 5.31).

	Unbalanced		Balanced	
	FA	NFA	FA	NFA
Featured Article	42	187	29	200
Non-Featured Article	1069	160130	28	201

Table 5.31.: Two Mode Network: Contingency Table on Betweenness and Redundancy

For the unbalanced data set there are more article classified correctly as for the balanced data set. For the balanced data set the number of featured and non-featured articles is roughly the same (see table 5.31).
Since the number of correctly classified featured articles is very low, also the precision and recall values are low and therefore also to a low CCI of 50.22% for the balanced data set. This might be caused due to the timespan on which the data set is built on. Only edits within a year are considered leading to a sparse network and therefore to a low betweenness. For a future work the whole history of edits might be taken into account.

	Unbalanced		Balanced	
	FA	NFA	FA	NFA
TP Rate	0.183	0.993	0.127	0.878
FP Rate	0.007	0.817	0.122	0.873
Precision	0.038	0.999	0.509	0.501
Recall	0.183	0.993	0.127	0.878
F1	0.063	0.996	0.203	0.638
ROC Area	0.711	0.711	0.517	0.517
CCI	99.2219%		50.2183%	

Table 5.32.: Two Mode Network: Correctness of Classifier on Betweenness and Redundancy

Classification with Betweenness, Redundancy and Path Length Since all other combinations of metrics, which are betweenness, redundancy and average path length, for the classifier did not work well, the last approach is to use all three metrics in order to classify articles. However, table 5.33 shows that only a few more article more were classified correctly.

	Unbalanced		Balanced	
	FA	NFA	FA	NFA
Featured Article	43	186	32	197
Non-Featured Article	1140	160059	31	198

Table 5.33.: Two Mode Network: Confusion Matrix on Betweenness, Redundancy and Path Length

For both the unbalanced and balanced data set, only a few more articles were classified correctly. For both data sets the classifier is biased towards non-featured articles.

Since the number of correctly classified articles for both the unbalanced and balanced data set is low, also the corresponding recalls, precisions and CCI are low. This was not unexpected, as all other combinations of metrics for the classifier, which were discussed before, produced roughly the same results.

	Unbalanced		Balanced	
	FA	NFA	FA	NFA
TP Rate	0.188	0.993	0.140	0.865
FP Rate	0.007	0.812	0.135	0.860
Precision	0.036	0.999	0.508	0.501
Recall	0.188	0.993	0.140	0.865
F1	0.060	0.996	0.220	0.635
ROC Area	0.741	0.741	0.548	0.548
CCI	99.1786%		50.2183%	

Table 5.34.: Two Mode Network: Accuracy on Betweenness, Redundancy and Path Length

Summary of Classification for Two-Mode Network To sum up there is no ideal classifier for the two-mode network with the given metrics of betweenness, path length and redundancy for both the unbalanced and balanced data set. A reason might be that the network consists only of edits within one year leading to a sparse network. Another reason might be that the given metrics should be replaced by others, that focus more on the two-mode nature and also take user into account. Therefore the fourth research question - whether featured articles can automatically be classified - must be negated for this two-mode network for classification.

6. Conclusion

This thesis tries to classify featured articles in the Norwegian Wikipedia and for some special parts also on the English Wikipedia. Therefore a data dump for both different language Wikipedias was downloaded. This data dump was then preprocessed in order to get rid of bots (only in the Norwegian data dump), redirects, disambiguations and articles that were shorter than 512 characters. After that four different networks were built on which different metrics were calculated.
The first network was article network, were two nodes are connected when there is a link from one article to the other article. For this network the average text length was calculated, which is a non network specific value. However the average text length for featured articles is about seventeen times higher than for non-featured articles. Also the average path length for featured articles in the article network is less than for non-featured articles. For the classification the metrics article length, local clustering coefficient and betweenness are used. All these metrics show different values for featured articles than for non-featured articles. Therefore the first research question - whether featured articles have different properties - can be affirmed that featured articles in article network do have different properties. Also the second research question - whether featured articles lie at structural holes - can be affirmed. Although the values for the local clustering coefficient are not as high as the ones Ingawale et al. received, however the average path length and the local clustering coefficient for featured articles is less than for non-featured articles, indicating structural holes. The third research question - whether featured articles can automatically be classified - can be confirmed for the balanced data set. The best results are achieved when using local clustering coefficient and average text length in a 10-fold cross validation with naive Bayes with a recall of 0.878, F1 of 0.151 and CCI of 98.6169% in the Norwegian Wikipedia and a recall of 0.472, F1 of 0.049 and CCI of 98.725% in the English Wikipedia for the unbalanced data set.

The same classifier returns a recall of 0.742, F1 of 0.805 and CCI of 82.0961% for the Norwegian Wikipedia and a recall of 0.839, F1 of 0.781 and 76.3827% for the English Wikipedia for the balanced data set, where the number of featured and non-featured articles is the same. The classification for the English Wikipedia did not perform well. Reasons might be the very sparse structure of the network. Also the CCI values for the unbalanced data set are not very meaningful as there are much more non-featured than featured articles, leading to a wrong percentage of correctly classified instances.
The second network constructed was the user network. Since this network only has users as nodes and edges are collaborated edits, there are no featured nodes. Instead the number of edges which come from a featured article are counted for each user. Then the betweenness is calculated and the correlation between those two metrics is plotted. The Spearman coefficient is 0.29 indicating a very weak correlation of these two metrics.
Ingawale et al. uses the collaboration network for their experiments. However, this network is not used in this thesis as it is very big even with a timespan of only one year of edits.
Instead the Two-Mode Network is used, which has actor nodes and user nodes in the same network. This thesis does not focus on the Two-Mode network. However, this thesis tries to find metrics that are suitable to answer the fourth research question - whether the analysis of the first three research questions can be extended to the Two-Mode network.This network has different network properties and therefore needs different metrics, for example it is not possible to calculate the local clustering coefficient as there is no node that is connected within the same set. Therefore it was necessary to find metrics that are suitable for this kind of network. For this network the redundancy, betweenness and average path length for every node is calculated. Instead of the local clustering coefficient the redundancy is calculated. The statistics for both values (see tables 5.14 and 5.15) show that there is only a very small difference for featured articles than for non-featured articles. If the margin of the differences is not taken into account then the second research question - whether featured articles lie at structural holes - can be affirmed. In summary featured articles do have different properties than non-featured articles and therefore the first research question - whether featured articles have different properties than non-featured articles - can also be affirmed. Featured articles show different degree distribution, article length and average path length for the article network, and degree distri-

bution, average path length and redundancy for the two-mode network. The last research question - whether featured articles can automatically be classified - must be negated for two-mode networks as the results for classification did not produce any relevant positive results. The best result with a CCI of 51.7467% for the balanced data set was achieved when using only the betweenness. This small CCI value means that the classifier is just slightly better than choosing by random. Therefore the assumption that similar metrics for the article network are also valid for the Two-Mode network must be negated.

6.1. Lessons Learned

Runtime Complexity and Complexity of collaboration networks The collaboration network with the articles being the nodes every two user working on the same article produces an edge, led to a rather dense graph with a density of $4,03\%$ compared to the graph with articles and link to other article as edges, which had a density of $0,023\%$. The maximum number of edges in a collaboration network can be calculated using $\frac{(n*n-1)}{2}$ where each node can be connected to all possible other nodes in the network but not to itself and because this graph is undirected that number can be divided the sum by 2. This formula shows that considering the Norwegian dataset of 165000 articles a high number of edges can be expected.
Combined with the expensive runtime complexity of the betweenness and clustering coefficient, experiments on collaboration graphs of Wikipedia tend to take a long time or require huge amounts of processing power.

Lack of resources For this thesis only a normal state of the art PC with a quad core CPU was available, but as the graphs described in this thesis tend to get really large, more computational power would be required to work with this amount of data. If access to a cluster would have been available, horizontal scalability would have been really important. The ability to distribute the workload to a big amount of nodes would have drastically reduced the time spent on working with the dataset. As a matter of fact,

working with graphs of this size on a single CPU borders to impracticability, as weeks are spent waiting for results.

In retrospective with the focus of horizontal scalability in mind, Apache spark running on a large scale Apache Hadoop cluster would be the way to go if such hardware would be available.

6.2. Future Work

This thesis uses Social Network Analysis in order to answer four research questions- whether featured articles have different properties than non-featured articles, whether they lie at structural holes, whether featured articles can automatically be classified and if these questions can also be answered for the two-mode network representation. Therefore four different networks are introduced, where only the article network had implicitly the whole history of edits in it. The other three network use only one year of edits made to the Norwegian Wikipedia. This might be the reason that the last research question must be neglected. As future work, a wider timespan might be used in order to receive better results. For this approach a better hardware is necessary in order to terminate in realistic time.
It might also be interesting looking at other metrics than presented in this thesis. Maybe those are more accurate in capturing the correlations between the position of a node in the network and the quality of its content. Also ones that are designed for two-mode networks, which also capture the interaction of user and articles.

List of Figures

List of Tables

List of Tables

Bibliography

Abdo, Alexandre H. and A. P. S. de Moura (2006). "Clustering as a measure of the local topology of networks." In: *Arxiv.Org*, pp. 1–5. arXiv: 0605235 [physics]. URL: http://arxiv.org/abs/physics/0605235v4http://arxiv.org/abs/physics/0605235 (cit. on p. 35).

Adler, B. Thomas et al. (2008). "Measuring author contributions to the Wikipedia." In: *Proceedings of the 4th International Symposium on Wikis - WikiSym '08*. New York, New York, USA: ACM Press, 15:1–15:10. ISBN: 9781605581286. DOI: 10.1145/1822258.1822279. URL: http://portal.acm.org/citation.cfm?doid=1822258.1822279 (cit. on p. 17).

Alexa. "Analysing Web-Traffic." In: URL: http://www.alexa.com/siteinfo/ (visited on 12/26/2015) (cit. on pp. 1, 9).

Barrat, a and M Weigt (2000). "On the properties of small-world network models." In: *European Physical Journal B* 13.3, pp. 547–560. ISSN: 14346028. DOI: 10.1007/s100510050067. arXiv: 9903411v2 [arXiv:cond-mat]. URL: <GotoISI>://WOS:000085407700019$\backslash$nhttp://www.springerlink.com/index/0HGUCD51T90CKB12.pdf (cit. on p. 34).

Berkeley. *GraphX*. URL: https://amplab.cs.berkeley.edu/wp-content/uploads/2013/05/grades-graphx_with_fonts.pdf (visited on 12/12/2015) (cit. on p. 100).

Blumenstock, Joshua E (2008). "Size matters: word count as a measure of quality on wikipedia." In: *Www*, pp. 1095–1096. ISSN: 08963207. DOI: 10.1145/1367497.1367673. URL: http://portal.acm.org/citation.cfm?id=1367673 (cit. on pp. 21, 55, 56, 73).

Borgatti, Steve (2010). "Strength of Weak Ties , Structural Holes , Closure and Small Worlds Strength of Weak Ties theory." In: (cit. on p. 28).

Brin, Sergey and Larry Page (2012). "Reprint of: The anatomy of a large-scale hypertextual web search engine." In: *Computer networks*. ISSN: 01697552. DOI: 10.1016/S0169-7552(98)00110-X. URL: http://www.sciencedirect.com/science/article/pii/S1389128612003611 (cit. on p. 19).

Buriol, L S et al. (2006). "Temporal analysis of the wikigraph." In: *Web Intelligence, 2006. WI 2006. IEEE/WIC/ACM International Conference on Web Intelligence*, pp. 45–51. DOI: 10.1109/WI.2006.164. URL: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4061340 (cit. on pp. 10, 21).

Burt, Ronald S. (2001). "Structural Holes versus Network Structure as Social Capital." In: *Social Capital: Theory and Research* May 2000, pp. 31–56. ISSN: 00018392. DOI: Burt{_}2001. URL: http://homes.chass.utoronto.ca/{~}wellman/gradnet05/burt-STRUCTURALHOLESvsNETWORKCLOSURE.pdf$\backslash$nhttp://aisel.aisnet.org/cgi/viewcontent.cgi?article=1188{\&}context=acis2005 (cit. on pp. 28, 29).

Csardi, Gabor and Tamas Nepusz. *igraph*. URL: http://igraph.org (cit. on p. 100).

Cygwin. *Cygwin*. URL: https://cygwin.com/index.html (cit. on p. 41).

Freeman, Linton C. (1977). "A Set of Measures of Centrality Based on Betweenness." In: *Sociometry* 40.No. 1, pp. 35–41 (cit. on p. 33).

Gao, Ge, Pamela Hinds, and Chen Zhao (2013). "Closure vs. Structural Holes: How Social Network Information and Culture Affect Choice of Collaborators." In: *Proceedings of the 2013 conference on Computer supported cooperative work - CSCW '13*. Ed. by Amy Bruckman et al. New York, New York, USA: ACM Press, p. 5. ISBN: 9781450313315. DOI: 10.1145/2441776.2441781. URL: http://dl.acm.org/citation.cfm?doid=2441776.2441781 (cit. on pp. 28, 29).

Goh, Ki Kwang-Il et al. (2002). "Classification of scale-free networks." In: *Proceedings of the National Academy of Sciences of the United States of America* 99.20, pp. 12583–12588. ISSN: 0027-8424. DOI: 10.1073/pnas.202301299. arXiv: 0205232 [cond-mat]. URL: http://www.pnas.org/cgi/content/long/99/20/12583$\backslash$nhttp://www.pnas.org/content/99/20/12583.short (cit. on p. 16).

Hagberg, Aric A., Daniel A. Schult, and Pieter J. Swart (2008). "Exploring network structure, dynamics, and function using NetworkX." In: *Pro-*

ceedings of the 7th Python in Science Conference (SciPy2008). Pasadena, CA USA, pp. 11–15 (cit. on p. 101).

Hall, Mark et al. (2009). "The WEKA Data Mining Software: An Update." In: *SIGKDD Explorations* 11 (1) (cit. on pp. 36, 37).

Hasan Dalip, Daniel et al. (2009). "Automatic quality assessment of content created collaboratively by web communities: a case study of wikipedia." In: *Proceedings of the 2009 joint international conference on Digital libraries - JCDL '09*. New York, New York, USA: ACM Press, pp. 295–304. ISBN: 9781605583228. DOI: 10.1145/1555400.1555449. URL: http://portal.acm.org/citation.cfm?doid=1555400.1555449 (cit. on p. 18).

HenkvD. *Number of articles on en.wikipedia.org and Gompertz extrapolation*. Own work. Licensed under CC BY-SA 3.0 via Commons. URL: https://commons.wikimedia.org/wiki/File:EnwikipediaGom.PNG#/media/File:EnwikipediaGom.PNG (visited on 01/03/2016) (cit. on p. 11).

Hu, Meiqun et al. (2007). "Measuring article quality in wikipedia." In: *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management - CIKM '07* April, p. 243. DOI: 10.1145/1321440.1321476. URL: http://portal.acm.org/citation.cfm?id=1321476$\backslash$nhttp://portal.acm.org/citation.cfm?doid=1321440.1321476 (cit. on p. 17).

Ingawale, Myshkin et al. (2013). "Network analysis of user generated content quality in Wikipedia." In: *Online Information Review* 37.4, pp. 602–619. ISSN: 1468-4527. DOI: 10.1108/OIR-03-2011-0182. URL: http://www.emeraldinsight.com/10.1108/OIR-03-2011-0182 (cit. on pp. iii, iv, 1–3, 20, 23, 24, 27, 28, 32–34, 63, 65, 69–71, 75, 84, 85).

Kamps, Jaap and Marijn Koolen (2008). "The Importance of Link Evidence in Wikipedia." In: *Advances in Information Retrieval*. Ed. by Craig Macdonald et al. Springer Berlin Heidelberg, pp. 270–282. ISBN: 978-3-540-78645-0. DOI: 10.1007/978-3-540-78646-7_26. URL: http://dx.doi.org/10.1007/978-3-540-78646-7_26 (cit. on p. 18).

— (2009). "Is Wikipedia link structure different?" In: *Proceedings of the Second ACM International Conference on Web Search and Data Mining - WSDM '09*, pp. 232–241. DOI: 10.1145/1498759.1498831. URL: http://portal.acm.org/citation.cfm?doid=1498759.1498831 (cit. on p. 19).

Laniado, David and Riccardo Tasso (2011). "Co-authorship 2.0: Patterns of collaboration in Wikipedia." In: *Proceedings of the 22nd ACM conference*

on Hypertext and hypermedia - HT '11. New York, New York, USA: ACM Press, pp. 201–210. ISBN: 9781450302562. DOI: 10.1145/1995966.1995994. URL: http://portal.acm.org/citation.cfm?doid=1995966.1995994 (cit. on pp. 16, 17, 20).

Latapy, Matthieu, Clémence Magnien, and Nathalie Del Vecchio (2008). "Basic notions for the analysis of large two-mode networks." In: *Social Networks* 30.1, pp. 31–48. ISSN: 03788733. DOI: 10.1016/j.socnet.2007.04.006. URL: http://linkinghub.elsevier.com/retrieve/pii/S0378873307000494 (cit. on p. 64).

Lex, Elisabeth et al. (2012). "Measuring the quality of web content using factual information." In: *Proceedings of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality - WebQuality '12* iii, p. 7. DOI: 10.1145/2184305.2184308. URL: http://dl.acm.org/citation.cfm?id=2184305.2184308 (cit. on p. 21).

Lih, Andrew (2004). "Wikipedia as Participatory Journalism : Reliable Sources ? Metrics for evaluating collaborative media as a news resource." In: *5th International Symposium on Online Journalism*, pp. 1–31. URL: http://www.ufrgs.br/limc/participativo/pdf/wikipedia.pdf (cit. on p. 20).

Lowd, Daniel and Pedro Domingos (2005). "Naive Bayes models for probability estimation." In: *22nd international conference on Machine learning*, pp. 529–536. DOI: 10.1145/1102351.1102418. URL: http://dl.acm.org/citation.cfm?id=1102351.1102418.

Newman, M. E. J. (2004). "Power laws, Pareto distributions and Zipf's law." In: 1. DOI: 10.1016/j.cities.2012.03.001. arXiv: 0412004 [cond-mat]. URL: http://arxiv.org/abs/cond-mat/0412004http://dx.doi.org/10.1016/j.cities.2012.03.001 (cit. on p. 19).

Peixoto, Tiago P. *Performance of graph-tool*. URL: https://graph-tool.skewed.de/performance (cit. on pp. 101, 102).

— (2014). "The graph-tool python library." In: *figshare*. DOI: 10.6084/m9.figshare.1164194. URL: http://figshare.com/articles/graph_tool/1164194 (visited on 09/10/2014) (cit. on pp. 98, 101).

Powers, David MW (2011). "Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation." In: *Journal of Machine Learning Technologies* 2.1, pp. 37 –63 (cit. on pp. 37, 38).

Qin, Xiangju and Pádraig Cunningham (2012). "Assessing the Quality of Wikipedia Pages Using Edit Longevity and Contributor Centrality." In:

CoRR abs/1206.2, p. 9. arXiv: 1206.2517. URL: http://arxiv.org/abs/1206.2517 (cit. on pp. 17, 18).

Royal, Cindy and D. Kapila (2008). "What's on Wikipedia, and What's Not . . . ?: Assessing Completeness of Information." In: *Social Science Computer Review* 27.1, pp. 138–148. ISSN: 0894-4393. DOI: 10.1177/0894439308321890. URL: http://ssc.sagepub.com/cgi/doi/10.1177/0894439308321890 (cit. on p. 10).

Spark, Apache. *Algorithms of GraphX*. URL: http://spark.apache.org/graphx/ (visited on 12/12/2015) (cit. on p. 100).

Stvilia, Besiki et al. (2005). "Assessing Inforamtion Quality of a Community - Based Encyclopedia." In: *Proceedings of the Tenth International Conference on Information Quality (ICIQ-05)*, pp. 442–454 (cit. on pp. 16, 20–22).

Suh, Bongwon et al. (2009). "The Singularity is Not Near: Slowing Growth of Wikipedia." In: *Proceedings of the 5th International Symposium on Wikis and Open Collaboration SE - WikiSym '09* 1.650, pp. 1–10. DOI: doi:10.1145/1641309.1641322. URL: citeulike-article-id:6405322$\backslash$nhttp://dx.doi.org/10.1145/1641309.1641322 (cit. on p. 10).

Viégas, Fernanda B, Martin Wattenberg, and Kushal Dave (2004). "Studying Cooperation and Conflict between Authors with history flow Visualizations." In: *Media* 6.1, pp. 575–582. DOI: http://doi.acm.org/10.1145/985692.985765. URL: http://alumni.media.mit.edu/~fviegas/papers/history_flow.pdf (cit. on p. 22).

Wasserman, Stanley and Katherine Faust (1994). *Social Network Analysis: Methods and Applications*. Vol. 1994, p. 825. ISBN: 0521387078. DOI: 10.1525/ae.1997.24.1.219 (cit. on pp. 20, 24–26, 28, 30, 31, 33).

Wikipedia, The Free Encyclopedia. *Autism*. URL: en.wikipedia.org/wiki/Autism (visited on 11/22/2015) (cit. on p. 7).

— *Help:Wiki markup*. URL: https://en.wikipedia.org/wiki/Help:Wiki_markup (visited on 01/04/2016) (cit. on p. 8).

— *List of Wikipedias*. URL: https://en.wikipedia.org/wiki/List_of_Wikipedias (visited on 12/26/2015) (cit. on p. 9).

— *MediaWiki*. URL: https://en.wikipedia.org/wiki/MediaWiki (visited on 02/12/2016) (cit. on p. 6).

— *Nupedia*. URL: en.wikipedia.org/wiki/Nupedia (visited on 10/20/2015) (cit. on pp. 5, 6).

— *Size of Wikipedia*. URL: https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia (visited on 12/17/2015).

Wikipedia, The Free Encyclopedia. *Utmerkede artikler*. URL: https://no.wikipedia.org/wiki/Wikipedia:Utmerkede_artikler (visited on 01/02/2016) (cit. on p. 14).

— *WikiDumps*. URL: https://dumps.wikimedia.org/enwiki/ (visited on 12/15/2015) (cit. on p. 39).

— *Wikipedia*. URL: https://no.wikipedia.org/wiki/Wikipedia (visited on 01/02/2016) (cit. on p. 11).

— *Wikipedia:Featured article criteria*. URL: https://en.wikipedia.org/wiki/Wikipedia:Featured_article_criteria (visited on 12/17/2015) (cit. on p. 15).

— *Wikipedia:Featured articles*. URL: https://en.wikipedia.org/wiki/Wikipedia:Featured_articles (visited on 02/12/2016) (cit. on p. 8).

— *Wikipedia:Modelling Wikipedia's growth*. URL: https://en.wikipedia.org/wiki/Wikipedia:Modelling_Wikipedia's_growth (visited on 01/03/2016) (cit. on p. 10).

Xiao, Wenjun et al. (2007). "Extended clustering coefficients:Generalization of clustering coefficients in small-world networks." In: *Journal of Systems Science and Systems Engineering* 16.3, pp. 370–382. ISSN: 1004-3756. DOI: 10.1007/s11518-007-5056-4. URL: <GotoISI>://WOS:000258569700008http://link.springer.com/10.1007/s11518-007-5056-4 (cit. on p. 35).

Zachte, Erich. *Wikimedia Statistics*. URL: https://stats.wikimedia.org/EN/Sitemap.htm (visited on 01/02/2016) (cit. on pp. 11–13).

Appendix A.

Evaluating and choosing a graph processing framework

After evaluating the amounts of data contained in the Wikipedia xml-dumps, several graph processing libraries and frameworks were compared and finally the python library *graph-tool* (Peixoto, 2014) was selected. The following sections document the decision for *graph-tool* and the factors involved in this decision.

Requirements

The chosen graph processing framework should be able to handle following issues:

- Small Memory footprint via efficient data structures
- Parallelization
- Algorithms implemented
- Data Import Capabilities
- Plotting Capabilities is seen as an advantage

These items will be discussed in detail in the following sections.

Small Memory footprint and efficient data structures As Wikipedia dumps consist of up to 4.5 million articles (in the English Wikipedia), which can basically be interpreted as graph nodes and up to 600 million links, the library or framework has to be able to handle large amounts of data efficiently. Being able to handle graphs of this size in memory is a big performance advantage. Therefore the library has to have a small memory footprint and use efficient data structures.

Parallelization The algorithms used by the graph processing library have to be able to be executed in parallel. The overhead of dividing the data into multiple units of work to be run by separate threads on separate processor cores should be minimal.

Algorithms Implemented It is seen as a big advantage if the required algorithms to calculate measurements and to analyse the graphs, are already been implemented. The required algorithms are:

- Calculation of the *in- and out-degrees*
- Calculation of the *betweenness values*
- Calculation of the *clustering coefficient*
- Calculation of the *largest component*

Data Import Capabilities It is to be expected that there will be no way to import the Wikipedia dumps directly into a graph-processing framework without prior converting and preprocessing of the dataset. Nevertheless a fast and simple way to add nodes and edges to the graph is required.

Plotting of Diagrams It is seen as an advantage that calculated data can be used to generate charts and diagrams. It should be easy to add additional information to the nodes like changing colour corresponding to its calculated measurement value.

Potential Candidates

The following libraries were evaluated according to the aspects discussed in the previous chapters. The libraries advantages and disadvantages will be summarized in this chapter.

Apache Spark GraphX GraphX started as a research project at the AMPLab from the UC Berkeley. It is built upon Apache Spark, and therefore uses Map Reduce to run algorithms distributed on a Hadoop/Spark Cluster. It can scale horizontally onto many cluster nodes, and because it supports Googles Pregel API, it is very flexible. GraphX is at the time still alpha and not suitable for production use. Apache Spark supports Python, Scala and Java, but GraphX only provides samples in Scala. The design of Spark and GraphX as a general parallel processing framework makes this library very flexible. It currently provides language bindings for Python, Scala, Java (*GraphX*).

Available algorithms As GraphX is currently in alpha state, it only provides some algorithms, but not all of the required algorithms are already implemented, and would have to be written. For example it does not provide algorithms for calculating clustering coefficients or the betweenness value.

According to the the GraphX Website it currently supports the following algorithms (*Algorithms of GraphX*):

- In Degrees
- Out Degrees
- Connected Components
- Strongly Connected Components

igraph

igraph is written in C/C++ and provides language bindings for C/C++, Python and R (*igraph*). There is not much information available on igraph,

except from its *git* repository and the "Getting started" guide. A quick review showed that it does not use *openMP* for parallelization, therefore it is fundamentally slower than graph libraries which support parallelisation. According to the benchmark it is approximately 2 times slower than graph-tool (*Performance of graph-tool*). The evaluation of igraph has also shown that there are not many examples of the usage of igraph, and also there is only a small documentation for this library available.

NetworkX

According to the API documentation NetworkX provides all required algorithms and can calculate the required coefficients. It is a python only module and therefore slowed down by the interpretative nature of python itself (Hagberg, Schult, and Swart, 2008). As the benchmark shows it is up to 20 times slower than for example graph-tool and iGraph (*Performance of graph-tool*).

Graph-tool

Graph-tool is a python module which provides extensive features for graph analysis. It currently support the following features (Peixoto, 2014):

- Directed and undirected Graphs
- Adding arbitrary information to the vertices and edges with property maps
- Plotting charts with dot cairo or graphviz
- Support for statistical measurements
- Support for centrality measures
- Support for clustering coefficients

Decision for Graph-Tool (Python)

After evaluating the libraries, it was decided to use the python *graph-tool* library. In this section the reason for this decision are summarized and discussed.

Advantages of Graph-tool

According to the benchmark it is faster than igraph or networkX (*Performance of graph-tool*). Graph-tool uses openMP in order to be able to run in parallel. It distributes the workload on all CPU cores of the host system, but it is not able to scale horizontally onto multiple hosts, for example to distribute work to multiple cluster nodes. Graph-tool is a python library and therefore allows for easy and fast prototyping, but as many parts of it are implemented in C++ it is still very fast. The evaluation of graph-tool shows, that many of the required algorithms are already implemented. It also has a good community support which could prove helpful during the development phase of the analysis tool.

Disadvantages of Graph-tool

As graph-tool has been developed primarily for python it is also the only supported language. This effectively limits the use of other programming languages. There are no other language bindings for graph-tool available. As the preprocessing and preparation of the dataset has already been developed in Java, this will lead to two different sets of tools. The optimizations in graph-tool, especially the use of external C++ code makes it difficult to debug exceptions and error conditions.

Argumentation for Graph-tool

Even with the limitations, which graph-tool imposes, it was still recognized as the best alternative to other available libraries and frameworks, because of

the generally good community support and the good coverage of the given requirements. And as at the time being, no access to large scale computing clusters is available, the necessity to scale vertically will be neglected.