Kevin Bassa, BSc

Validation of Information: On-the-Fly Data Set Generation for Single Fact Validation

MASTER'S THESIS

to achieve the university degree of
Diplom-Ingenieur

Master's degree programme: Software Engineering and Management

submitted to

Graz University of Technology

Supervisor

Dr. Roman Kern

Knowledge Technologies Institute

Advisor
Dr. Mark Kröll

Graz, August 2016

AFFIDAVIT

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 indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present master‘s thesis.

Date

Signature

Abstract

Information validation is the process of determining whether a certain piece of information is true or false. Existing research in this area focuses on specific domains, but neglects cross-domain relations. This work will attempt to fill this gap and examine how various domains deal with the validation of information, providing a big picture across multiple domains. Therefore, we study how research areas, application domains and their definition of related terms in the field of information validation are related to each other, and show that there is no uniform use of the key terms. In addition we give an overview of existing fact finding approaches, with a focus on the data sets used for evaluation. We show that even baseline methods already achieve very good results, and that more sophisticated methods often improve the results only when they are tailored to specific data sets. Finally, we present the first step towards a new dynamic approach for information validation, which will generate a data set for existing fact finding methods on the fly by utilizing web search engines and information extraction tools. We show that with some limitations, it is possible to use existing fact finding methods to validate facts without a preexisting data set. We generate four different data sets with this approach, and use them to compare seven existing fact finding methods to each other. We discover that the performance of the fact validation process is strongly dependent on the type of fact that has to be validated as well as on the quality of the used information extraction tool.

Validierung von Information ist der Prozess zu bestimmen, ob eine bestimmte Information wahr oder falsch ist. Bestehende Forschung in diesem Gebiet richtet sich dabei hauptsächlich auf einzelne Bereiche und verabsäumt es, Beziehungen über mehrere Bereiche zu behandeln. Diese Arbeit wird diese Lücke füllen und untersuchen, wie verschiedene Bereiche mit der Validierung von Information umgehen. Dafür werden Forschungsgebiete, Anwendungsbereiche und deren Verwendung von ähnlichen Ter-

men analysiert, wobei sich herausstellt, dass es keine einheitliche Verwendung der Terme gibt. Als Nächstes werden bestehende Methoden zur Validierung von Fakten verglichen, mit einem Fokus auf die dafür verwendeten Datensätze. Wir zeigen, dass schon die Basisalgorithmen sehr gute Ergebnisse erzielen, und komplexere Algorithmen oft nur dann deutlich besser sind, wenn sie auf bestimmte Datensätze abgestimmt sind. Zum Abschluss wird ein erster Schritt zu einem neuen, dynamischen Ansatz zur Validierung von Fakten präsentiert, bei dem ein Datensatz mithilfe von Suchmaschinen- und Informationsextraktionstools dynamisch generiert wird. Wir zeigen, dass es mit ein paar Einschränkungen möglich ist, bestehende Methoden zur Validierung von Fakten ohne einem zuvor existierenden Datensatz zu nutzen. Der präsentierte Ansatz wird des Weiteren dazu genutzt, um vier Datensätze zu generieren, und damit sieben Methoden zur Validierung von Fakten miteinander zu vergleichen. Dabei zeigt sich, dass die Leistung des Prozesses zur Validierung von Fakten stark von der Art des Faktes, der validiert werden soll, und von der Qualität des Informationsextraktionstools, abhängt.

Contents

List of Figures

List of Tables

1 Introduction

On the web, massive amounts of information are available, ever-growing since the web's commercial release in the early 90s. This growth was driven by easy extensibility and accessibility, which also caused one disadvantage – the spreading of erroneous or fake contents. Wrong (or conflicting) information makes it hard for users to distinguish between what is true and what is not. Conflicting information is most commonplace in user-generated content, for example, in debates on social networks or entries in online encyclopedias, where information is sometimes only little or not at all verified for its truthfulness. But even non-user-generated content, as for example news stories published by online news provider or information about objects sold in electronic commerce can be erroneous. There are several problems with this situation. Not only are single users unsure about whom to trust on the internet, but there are also many projects exploiting the wealth of freely and easily accessible information on the web by automatically acquiring knowledge and using it for internal decision making processes, without knowing how valid the acquired knowledge actually is.

Since the public became aware of this problem, plenty of work has been done in this field of research, spreading over many different domains. Sometimes, those works have a very similar goal, but are conducted unrelated to each other. Along with the different domains, there is another reason for this. For the task of assessing the credibility of information, numerous different terms with different established definitions have been used. There is no uniform use of these terms. This makes it hard to grasp the meta-concept of the field of information validation, since so much research is done independently from others, without showing the actual relations that exist between different works. The goal of the thesis is to close this gap and create a structured overview of where, why and what has been done in the area of information validation.

The first part of the thesis (Chapter 2) will reveal connections across the whole field of information validation, starting with the analysis of all different terms that are being used, with the goal to create an understanding of their relation to each other. We show that there is no uniform use of the key terms. Moreover, not only are different terms used in the same context, but their definitions also vary. Secondly, we give an overview of research areas working on this topic and application domains benefiting from it, which will show how important information validation is for each of the different domains and what similarities exist among them. We identified five main research areas, with fact finding being the most relevant one. The two most important application domains that have been found are social media and news. This comes from the circumstances that these two domains have a massive amount of publicly available, often user-made content, which tends to be error prone. At the end, a list of languages which have been used will be given, stating what the differences are and why they have been used. Here, it will become apparent that English is by far the most used one.

In the second part of the thesis (Chapter 3), we will tackle approaches to assess the validity of information, with a main focus on various types of approaches and their used data sets. The chapter will be divided into two parts, content-based and meta-information-based assessment of information validity. We will show what the main differences between the approaches are, and how well these approaches have been evaluated. As it turns out, for content-based approaches, even the baseline methods already achieve very good results, and more sophisticated methods often improve the results only when they are tailored to specific data sets. In contrast, meta-information-based approaches have not been compared to each other at all, as they all use their own data set.

The last part of the thesis (Chapter 4) includes the implementation of the first step towards a new dynamic approach for information validation. The final goal of the dynamic approach is to use existing fact finding algorithms for the validation of single facts. This is accomplished by utilizing web search engines and information extraction tools to generate a data set on-the-fly. For this purpose, all freely available web search engine APIs have been tested, which turned out to be just two (Google and Bing). As these two limit their users to a specific amount of queries and returned results, a combination of both is used to maximize the amount of returned result websites. For

the information extraction step, seven state-of-the-art information extraction systems have been tested. The need for a uniform output of facts brought us to use the entity-relation extraction tool MitIE for this step, which is limited to a set of predefined relation models. These constraints led to a dynamic approach with some limitations, but we have still proven that it is possible to use existing fact finding methods to validate facts without a preexisting data set. We used this approach to generate four different data sets and compare seven existing fact finding methods to each other. We discovered that the outcome of the fact validation process is strongly dependent on the type of fact that has to be validated, and on the quality of the applied information extraction tool.

2 Theoretical Background

2.1 Literature Search Pattern

As the scope of the master thesis is very comprehensive, a systematic literature search was needed which would maximize the chance of finding as much related work as possible. To achieve this goal, we have used three different search engines / scientific libraries. In each of them, a combination of various search terms was used, and the results thereof were checked by reading the abstracts. All papers that we considered to be relevant were stored for further use. In the second step, we retrieved relevant papers which were cited by or cited one of the existing papers, which expands the search to papers which do not contain any of the used search terms. The details are listed below.

1. Used search terms:
   - (('believability' $\oplus$ 'certainty' $\oplus$ 'correctness' $\oplus$ 'credibility' $\oplus$ 'fidelity' $\oplus$ 'reliability' $\oplus$ 'trustworthiness' $\oplus$ 'trust' $\oplus$ 'truthfulness' $\oplus$ 'truth' $\oplus$ 'validity' $\oplus$ 'Veracity') $\wedge$ ((information $\oplus$ data)$\vee$(assessment $\oplus$ assess)))
   - information quality
   - open information extraction
   - ((truth $\oplus$ knowledge $\oplus$ fact) $\vee$ (finding $\oplus$ discovery)
2. Subsequent sources:
   - For most relevant papers: citing + cited by
3. Used search engine / scientific libraries:
   - http://dl.acm.org/
   - https://www.ieee.org/index.html

- https://scholar.google.at/

Before using this systematic approach, a normal, manual literature search with some basic terms like 'information validation' has been conducted. After a short period of time it became apparent that no uniform definition of this topic existed. Works in different research areas and application domains used various terms to describe the same or similar tasks. This led to a collection of many different terms. We then used these terms, as mentioned above, for a systematic literature search. After the first round of literature search, where the abstracts of promising papers had been checked, 81 papers related to information validation were found. In the second round, this list increased to a total of 155 papers by additionally using source citations. To ensure a basic overview we created a mind map (see appendices Figure .1) which roughly linked all papers to specific topics. Although not all papers were useful, all papers were somehow related to information validation. We used these papers as basis for Chapter 2. In this phase, all read papers where additionally rated by their quality and usefulness for this thesis. The criterion for a paper to be rated very good was that it had to contain a detailed description of any information validation assessment together with at least a simple evaluation, or to compare multiple existing approaches for this task. In the third round, we took these papers for the creation of a new mind map (see appendices Figure .2) with a new structure linking the papers to their precise topics. These were a total of 42 papers and became the basis for Chapter 3.

2.2 Definitions of Related Terms

2.2.1 Overview

During the research we have found nine terms, which had an equal or at least very similar meaning to validity and were used at least once in the area of validation of information:

- Believability
- Certainty

- Correctness
- Credibility
- Fidelity
- Reliability
- Trustworthiness
- Truthfulness
- Validity (root)
- Veracity

To give the reader a short overview, in the following subsections we will describe how and where each of the terms was used. The literature sources for this overview included all papers which were perceived as related to the topic. They were manually derived with the search pattern listed above, and should cover all important domains occurring in or overlapping with the validation of information domain.

2.2.2 Believability

Definition by [30]: "To have confidence in the truth, the existence, or the reliability of something, although without absolute proof that one is right in doing so".

Believability is frequently used in the information quality domain. It is one of the quality dimensions consistently mentioned in papers on information quality [6, 16, 20, 21, 25, 79, 82, 102, 118, 120, 122, 144]. These papers offer 3 definitions of believability. [115] definition of believability is "the extent to which information is regarded as true and credible", [65] also termed believability as "trustworthiness" and at last, [144] defined believability "as the degree to which the information is accepted to be correct, true, real and credible". Further connections are found in [7, 24, 73] - papers on credibility analysis of micro blogs, where they all use believability in their definition of the word credibility.

Summary: The term believability is mainly used in the information quality domain, as it is one of the quality dimensions. It has a strong link to trustworthiness and credibility.

2.2.3 Certainty

Definition by [30]: "The state of being certain (free from doubt or reservation; confident; sure)".

Certainty (or likewise 'uncertainty') is also mainly used in the fact finding domain. The term is rarely used directly like in 'certainty of information', but often in specific intermediate steps to calculate the validity of information, where a frequently used term is 'data certainty', as for example in [32, 38, 110, 113, 143]. [91] link certainty to veracity by defining that "data veracity includes two aspects, data certainty and data trustworthiness" and certainty to reliability by defining data certainty "by statistical reliability of data". Other occurrences are in the data quality domain, where certainty is also sometimes but not always used for specific metrics for one of the dimensions [13, 59].

Summary: The term certainty finds nearly no use at all as part of the general phrase 'certainty of information'. Instead, it is often used in connection with specific mathematical calculations which are used in the validation of information domain.

2.2.4 Correctness

Definition by [30]: "Conforming to fact or truth; free from error; accurate".

In truth finding, the term correctness is used in the context of 'correctness of information', where information refers to the input, hence a fact [10, 36, 38, 48, 83, 88, 89, 142, 143, 147]. The use in the question answering domain, with 'correctness of the answer', is reasonably the same, with the difference that an answer can consist of one or more facts and its correctness depends on the question, too [51, 105]. Correctness is also used in the information quality domain, either as quality dimension [16] or in the description of an indicator of a quality dimension [84, 145]. In [114], semantic correctness as part of accuracy is described as "degree of correctness and validity of the data in comparison to the real world or with the reference data agreed to be correct." Correctness is also used in the area of information extraction, in

the context of correctness of extracted facts, for which a confidence score is calculated [14, 39, 40].

Summary: The term correctness is a term used often in all domains in the context of 'correctness of information'. Its main use is in the truth finding domain.

2.2.5 Credibility

Definition by [30]: "The quality of being believable or worthy of trust".

Credibility is most often used in connection to the term information, namely 'information credibility', whereby the term always refers to a specific source. Sources are arbitrary websites [11, 67, 75, 77, 123], articles in encyclopedias[4, 93], posts in blogs [70, 128], reviews [80] or posts on social networks [5, 7, 24, 63, 73]. In all these papers credibility is always used in the sense of trustworthiness, as for example in [93], which says that the main credibility aspect of wikipedia is citing and referencing external sources that are credible, verifiable, and trusted. Credibility is defined in [62] as the ability to inspire belief or trust and as information accuracy and veracity. It is also defined as "trustworthiness, believability, reliability, accuracy, fairness, objectivity, and dozens of other concepts and combination thereof" [45] Information credibility is also used as single metric for the dimension believability in [144], where it is used together with reliability and calculated by the "use of trust annotations made by several individuals to derive an assessment of the sources' reliability and credibility." Furthermore, trustworthiness and believability are mentioned as synonyms for credibility in [17]: "credibility is related to trustworthiness of the data set, as well as other quality dimensions such as provenance, verifiability, believability, and licensing." This supports the universal use of credibility. According to [31], trustworthiness, believability, validity and reliability are synonyms for credibility, whereas veracity additionally links to credibility. This makes credibility one of the most expressive terms language wise.

Summary: The term credibility is mainly used in connection with information credibility, and has been used in papers concerning the credibility of web-

sites, blogs posts, social network posts, reviews or articles in encyclopedias. In the reviewed literature, it is most related to the term believability.

2.2.6 Fidelity

Definition by [30]: "Adherence to fact or detail".

Fidelity is an extremely rarely used word in the field of information validation. In all studied papers, it only occurs once in [94], as a word which is used in relation to the veracity. The meaning of the word in English would be rather fitting, as it is, according to [31], a synonym of 'trustworthiness' and 'veracity', and links itself to 'reliability'. But since other terms are simply more spot-on, and fidelity does not provide any useful additional meaning, it is just left out by the information validation community.

Summary: The term fidelity can be neglected, as it is never used as main term in the examined papers, but only once to describe the term veracity.

2.2.7 Reliability

Definition by [30]: "The ability to be relied on or depended on, as for accuracy, honesty, or achievement".

Reliability often occurs in the data quality domain, as it is again, similar to believability, repeatedly mentioned as one of the dimension of data quality [16, 79, 82, 84, 118, 120]. Its definition can vary as it depends on the application. Sometimes it is used as a main dimension, consisting of multiple sub-dimensions as in [84], sometimes it occurs only as a metric of a dimension as in [145]. In other works regarding the credibility of twitter posts [63], open source information [94] or truth finding [10, 86, 129], the reliability refers to the author or source in the sense of trustworthiness, while another term is reserved for the information itself (often used: credibility). Keep in mind that eventhough applicable to the majority of works, this does not make it the only truth, as there are still many others who use these terms in different senses.

Summary: The term reliability is very common in various domains. In most cases it is used in connection to sources.

2.2.8 Trustworthiness

Definition by [30]: "Deserving of trust or confidence; dependable; reliable;".

[144] worked on information quality assessment for linked open data, where various different existing approaches have been studied and a core set of twenty-three data quality dimensions have been extracted, which were split into six main groups. One of them was trustworthiness. Part of the group trustworthiness were the four dimensions reputation, believability, verifiability and objectivity, which again used, among many others, credibility, reliability and trustworthiness as metric. [46] use provenance to calculate the trustworthiness and thus also the quality of linked open data, while [16] uses reputation. This examples shows that in the information quality domain, trustworthiness is an expressive term that is connected with many similar words, and that its exact definition can vary. In the truth finding domain, trustworthiness is the most used word. There are many different approaches for truth finding, but trustworthiness is always used as the main term (among all available words), and it is always used with the same meaning [48, 49, 56, 83, 86, 89, 95, 110, 111, 126, 142, 143]. [113] describe the execution of a fact-finder algorithm with "[they] iteratively calculate the trustworthiness of each source given the belief in its claims, and the belief in each claim given the trustworthiness of its sources", which shows the link between trustworthiness and believability. [48] also has a recursive definition, but uses correctness instead of believability:"A correct answer is returned by many trusted views and a trustworthy view returns many correct answers." Side note: And although the domain is called truth finding, the term truthfulness is hardly ever used. Additionally to these two big domains, trustworthiness is also used in various other areas, for example in assessing content validity in social media [66], assessing trustworthiness of location data [28] or in web search and information credibility analysis[123], where trustworthiness is used together with many other terms, namely validity, credibility or correctness of information.

Summary: The term trustworthiness is one of the most used terms in the whole validation of information domain, and also the main term in the sub-domain truth finding.

2.2.9 Truthfulness

Definition by [30]: "Conforming to truth (conformity with fact or reality; verity)".

Truthfulness is a sparsely utilized term. In the question answering domain, [96] compares the problem of validating an answer to estimating the truthfulness of the associated validation statement, which makes these two terms synonyms.

In the truth finding domain, truthfulness is typically used with regard to the truthfulness of the answer, and therefore, the most probable truth [10, 36, 88, 89]. Similar examples can be found in [5, 11, 66], where the truthfulness of statements or Twitter Tweets, respectively, is calculated.

Summary: The term truthfulness is only sparsely used, most often still in the truth finding domain.

2.2.10 Validity

Definition by [30]: "The state or quality of being valid (sound,just, well-founded)".

The term "validity" is most often used in connection with the information quality domain. Validity is often mentioned as one of the dimensions of information quality, which leads to an overlap of the information quality domain with the information validation domain. Therefore, the term occurs in papers which either deal with the theoretical definition [82, 120] or with the assessment of information quality [3, 16, 17, 20, 46, 59, 60, 122, 135, 144, 145]. Validity, which is also used in metrics for information quality, is a very flexible term. [46] describes it as "one of the most flexible and important metrics, because it encodes context- and application-specific requirements".

Depending on the use of the term, its definition can overlap with the trustworthiness, as for instance in [135], but it is often defined in a specific way, as only a part of general meaning of trustworthiness, which rather makes it a sub-category. An example is the definition "Validity requires that the data set conforms to a set of custom constraints, expressed as logical rules" in [46]. In the truth finding domain, the term is once used like a substitute for correctness [83]. It also has a very similar meaning in the question answering domain [96, 97, 106], where a validity-score is calculated for each found answer. Here, the term would be again replaceable with correctness or truthfulness. [66] uses the metrics contributor, content and context validity to calculate the truthfulness and trustworthiness of social media posts, which again supports the close link between these terms.

Summary: The term validity finds its main use in the information quality domain. It often refers to the information itself, for example the validity of a fact, where often used synonyms are truthfulness or correctness. In all other domains it is used - although rarely - in a more global way, for example in 'the validity of blog posts'. In this context, similar used terms are trustworthiness or credibility.

2.2.11 Veracity

Definition by [30]: "Conformity to truth or fact; accuracy".

[117] mentions that veracity is the forth 'v' of the three 'v' features of big data: volume, velocity and variety. Since truth finding is increasingly important in the big data area, it is consequential that the verbal fitting term veracity is also increasingly used in the general truth finding domain, and therefore, also in the information validation domain. An example of the universality of the term is given with [94] work about automatic veracity assessment of open source information, where the author states "In this paper, we use Veracity interchangeably with trust, reliability and credibility." [18] says "Typically, the goal of truth finding is to determine the veracity of multi-source, conflicting data and return, as outputs, a veracity label and a confidence score for each data value, along with the trustworthiness score of each source claiming it", which reinforces its universality. But

nevertheless, the term veracity is not very common, which is reflected by other scientific works in the truth finding domain which use the term veracity only once, choosing to use the other terms as the main term instead, like trustworthiness or believability [56, 110, 111, 113, 129, 142]. Language wise this again matches the synonyms of veracity, as it is most similar to trustworthiness and truthfulness.

Summary: The term veracity is used almost exclusively in the truth finding domain. The most used synonym is trustworthiness.

2.3 Relations between Terms

In the following subsections we show the relations between the previously described terms. This is done by first exploiting the synonym relations provided by the online thesaurus [31], and then by using the existing papers to compare which terms got used to describe each other and how often a term occurred overall. For this meta analysis we used the 155 papers related to information validation, as described in Section 2.1.

2.3.1 Thesaurus Synonyms

To receive a language overview of the terms, we have checked their relations in the online thesaurus [31]. Although thesauri in general may not provide a complete list of synonyms, they are adequate for a providing a simple overview. In Figure 2.1, the relations between the terms in [31] are displayed as a graph. One can easily recognize that the graph mostly (but not always) coincides with the actual use of the words, as described in the sections above and shown in the Table 2.1, with trustworthiness, credibility and veracity being the most connected and centred terms, and only veracity being a less used term.

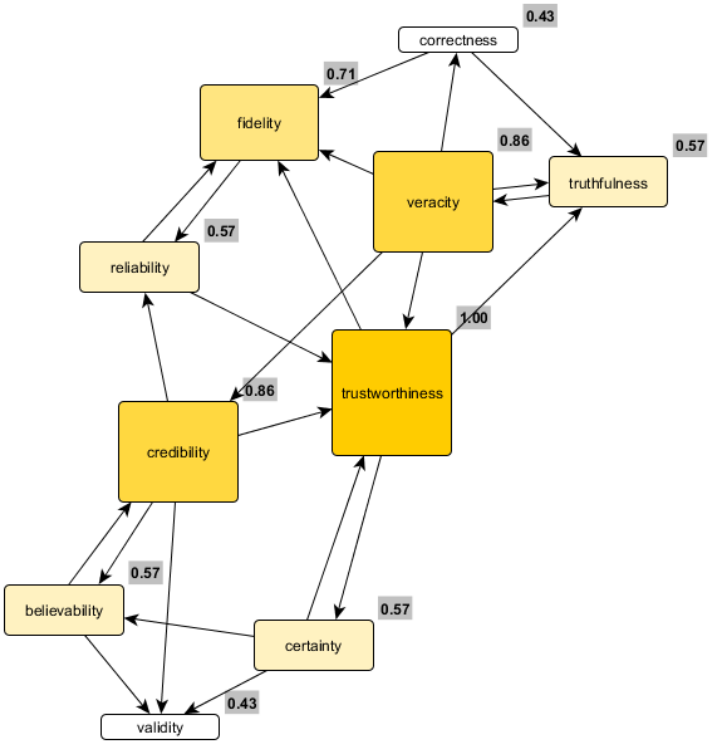

Figure 2.1: Links between synonyms as on Thesaurus.com, displayed as graph

2.3.2 Used Relations between Terms

Table 2.1 shows when terms got used to describe another term. In comparison to all examined papers, only very few actually defined or at least described one of the terms. These relations have been described in more detail in the above sections 2.2.2 - 2.2.11.

	Believability	Certainty	Correctness	Credibility	Fidelity	Reliability	Trustworthiness	Truthfulness	Validity	Veracity
Believability	-			[7, 24, 73, 115]			[65, 113]			
Certainty		-				[91]				[91]
Correctness			-						[114]	
Credibility	[7, 17, 62, 73, 144]			-		[7]	[7, 17, 62, 93]			
Fidelity					-					[94]
Reliability						-				
Trustworthiness	[57, 113]		[48]				-			
Truthfulness								-	[96]	
Validity				[103]		[103]	[66, 135]	[66]	-	
Veracity				[94]	[94]	[94]	[18, 94, 142]	[66]		-

Table 2.1: Used relations between terms in papers

2.3.3 Total Word Count

Table 2.2 shows how often a term occured in the examined papers. As can be seen in the first part of the table, the most used term is credibility, followed by reliability and trustworthiness. The second part displays the numbers for a more generous search which does not only include the noun, but also the adjective of a term (for example, the query 'credib' would match for both 'Credibility' and 'credible', and also includes other forms like 'credibly'). In addition, this part has 'correct' and 'truth' under the top used terms. These numbers do not directly reflect their use for information validation, as every occurrence of the term has been counted, including occurrences in other topics that are dealt with in the examined papers. Nevertheless, they show an approximate trend for each term.

Term 1	Count 1	Term 2	Count 2
Credibility	1364	trust	3100
Reliability	978	correct	1945
Trustworthiness	609	credib	1676
Correctness	290	truth	1403
Veracity	221	reliab	1289
Certainty	205	valid	860
Validity	198	certain	838
Truthfulness	137	belie	604
Believability	117	veraci	235
Fidelity	4	fidel	5

Table 2.2: Overview of how often a term occurs in papers related to information validation

2.4 Research Areas

In the field of NLP we have found five research areas which deal with information validity. Some of them overlap, and some of them are independent of each other, but all of them contribute to the topic of information validity:

- Information Quality
- Fact Finding
- Question Answering
- Information Extraction
- Credibility Assessment

2.4.1 Information Quality

2.4.1.1 General

The information quality domain is quite old, since it already had fields of application before the era of the internet. An example is [64], who wrote a dissertation about quality control of information in databases and management information systems in the year of 1972. Another term which is sometimes used instead of information quality is data quality. [30] states: "Data itself has no meaning, but becomes information when it is interpreted. Information is a collection of facts or data that is communicated. However, in many contexts they are considered and are used as synonyms." In our case it is always used as a synonym. A commonly used definition for information quality is "fitness for use", which was originally defined in 1974 by [71]. Because of the rapidly increasing amount of information available on the internet, the importance of being able to assess the information quality is rising as well, which is being reflected by the continually released research work in this domain.

2.4.1.2 Information Validation in Information Quality

Various dimensions which can then be treated independently from each other have been defined to determine the quality of information. How many and which dimensions are used depends on a variety of existing definitions. The important part for this work are the dimensions concerning the validity of information. These are:

- Believability
- Correctness
- Credibility
- Reliability
- Trustworthiness
- Validity

[82], [79] and [16] compared quality dimensions used in previous works in the information quality domain, which leads to the six dimensions listed

above. They are used in different works and describe the validity aspect of information. The exact definitions of the terms is described in section 2.2. For each dimension, assessable metrics are defined. Depending on the field of application, the amount and scope of these metrics are often limited and only give an insufficient assessment of the dimension. That is because of the fact that most of the dimensions were defined with only theoretical background in mind, independent from the actual use.

2.4.2 Fact Finding

2.4.2.1 General

[142] defines fact finding as "The goal of fact finding is to identify if a fact for an object is true, whereby the object may have multiple conflicting facts". We only use the term fact finding in connection with the NLP domain, where it is also sometimes called fact checking, truth finding or truth discovery. It is quite a new area, with increasing research activities in the past 9 years. The increase in interest in this area is coming from the massive growth of data available through the internet. The fact finding domain emerged from the question answering domain, and started to be used independently around 2007, with [136] and [142]. At the same time, the area of data fusion, which makes use of fact finding, too, also expanded to NLP such as [22] or [108]. Data fusion comes from the area of data integration, which main use was in merging databases. Bear in mind that all these terms might have been used much earlier, but not in connection to facts in natural language processing.

2.4.2.2 Information Validation in Fact Finding

The fact finding domain is entirely included in the topic of information validation. It covers the validation of single facts with various approaches and multiple specific algorithms.

2.4.3 Question Answering

2.4.3.1 General

Questions answering (short QA) refers to building a system which answers questions asked by humans. [97] describes it with "question answering systems search for answers to a natural language question either on the Web or in a local document collection." This discipline can be divided into two subgroups, i.e. open-domain QA and closed-domain QA. The difference between these two subgroups is, as the name suggests, that closed-domain QA is restricted to answering questions of a specific domain while open-domain QA is domain independent. The first steps in closed-domain QA were made with LUNAR in 1971, a system which was limited to questions about geological analysis of rocks and had a knowledge base hand-written by experts of the domain [72]. Later, much more sophisticated open-domain QA systems received more and more attention, with WATSON [44] being one of the most popular ones. The basic steps for a question answering system are:

- **Query building:** Decompose the question into its parts and build several queries with them.
- **Answer Extraction:** Use the queries to search the knowledge base for all possible answers and extract them.
- **Answer Validation:** Compute a confidence score for each extracted answer.

2.4.3.2 Information Validation in Question Answering

From the above listed three basic steps of question answering, the last one is the one overlapping with the information validation domain. According to [97], the goal of answer validation is "filtering out improper candidates by checking how adequate a candidate answer is with respect to a given question." In contrast to information validation, the task of answer validation is more complicated, because in addition to validating a single fact, the relevance of the answer to the given question has to be computed. But in this thesis, when referring to the question answering domain, we refer

mainly to the validation of single facts. For this validation, mainly fast and simple approaches (for example VOTE) are utilized, as the time required for answering a question is critical [2, 92, 97].

2.4.4 Information Extraction

2.4.4.1 General

[40] describes information extraction (IE) as "venerable technology that maps natural-language text into structured relational data." Furthermore, the authors say "At the core of an IE system is an extractor, which processes text; it overlooks irrelevant words and phrases and attempts to home in on entities and the relationships between them." There are two types of IE systems, domain-specific and open IE systems. Domain-specific IE systems were the first ones to be developed, as the task is much simpler in a small, specified environment. [40] also gives a good overview of the methods used for IE systems:

- **Knowledge-Based Methods** relied on some form of pattern-matching rules that were crafted manually for each domain. These systems were clearly not scalable or portable across domains.
- **Supervised Methods** are used by modern IE. They automatically learn an extractor from a training set in which domain-specific examples have been tagged. With this machine-learning approach, an IE system uses a domain-independent architecture and sentence analyzer. The development of suitable training data for IE requires substantial effort and expertise. The amount of manual effort scales linearly with the number of relations of interest, and these target relations must be specified in advance.
- **Self-Supervised Methods** are the most recent development in the IE domain. Here, IE systems are automated by learning to label their own training examples using only a small set of domain-independent extraction patterns.

2.4.4.2 Information Validation in Information Extraction

One part of information extraction is to assess whether the extracted information is correct. This is important to improve the precision of IE. There are two different kinds of assessment. The first one only uses all the extracted information and tries to find out which information is most relevant and correct. The second one additionally verifies the correctness of the extracted information using the web (or any other knowledge base independent of the original source). Redundancy based approaches are often used in this step, as for example in [12, 39, 41], which is closely related to the approaches used in the fact finding domain. For the first step, multiple other approaches have been suggested, mainly driven by the research done in the open IE domain, where facts tend to be more difficult to extract than in domain-specific IE and thus need to be validated more accurately. An example of this is [42], which uses 19 features for a logistic regression classifier to assign a confidence score to each extraction, which is also done by [100], but with different features. Another different approach is explained by [141], which assigns the validity score to a pattern depending on the semantic similarity of the attributes of its extracted facts and then subsequently propagates this score to all facts extracted by this pattern. For this thesis, only the second step (determining if a piece of information is true or false) is relevant, and thus the first one will be neglected. But it can still be seen that research in information extraction, especially open information extraction, has created some interesting approaches, which will be further expanded upon in Chapter 3.1.

2.4.5 Credibility Assessment

2.4.5.1 General

The research area 'credibility assessment' comprises of the work which deals with the credibility of more complex objects in a specific context, which usually provide additional meta information. These 'objects' mainly are:

- Encyclopaedia articles and their authors
- Forum posts

- Blogs
- Social media posts
- Arbitrary websites
- User reviews
- News portals and articles

Work in this domain also started in the mid-2000s, with the same underlying cause as fact finding: The amount of information available on the internet started to get enormous, and thus the need to detect true information and sources which provide true information has arised. [134].

2.4.5.2 Information Validation in Credibility Assessment

This domain is a direct and important part for the general area of 'information validation'. In contrast to content-based information validation, e.g. in the fact finding domain, the information validation in this domain is mainly meta-information-based. Naturally, this meta-information is different for each subject, but some typical examples are timestamps, the length of a text or the number of facts.

2.5 Relations between Terms and Research Areas

Table 2.2 shows which terms are commonly used in which research areas. In each column, the frequency of the terms is color-coded, starting with white for the least used term and ending at dark green for the most used term. A word counts as 'commonly used' if it occurs at least three times in a paper. We have chosen the number 3 because many of the words on the list would be mentioned at least once in the related work section of a paper, without the paper itself actually using the word. To acquire these data, we first divided the papers into five research domains and then parsed them for the terms automatically. It can be seen, that work regarding general credibility assessment mainly uses the term credibility and trustworthiness, while in fact finding, trustworthiness and correctness are the top two terms. In information quality, a wide variety of terms is

used, with reliability, believability, validity and correctness all being used frequently. On the contrary, work done in the question answering domain has a very restricted use of only three terms, with 'correctness' being the one most used. Information extraction does not really use any of these terms multiple times at all, with only 'correctness' being mentioned more often in one single paper. This is because only a short part of the work dealing with information extraction deals with credibility of the extracted information, and any definitions are neglected there, which leads to a restricted use of the word 'correct'.

	Information Quality	Fact Finding	Question Answering	Information Extraction	General Credibility Assessment	Sum
Believability	9	1			1	11
Certainty	2	5			5	12
Correctness	7	8	6	1	2	24
Credability	5				23	28
Fidelity						0
Reliability	10	3	3		7	23
Trustworthiness	4	14			13	31
Truthfulness		3			2	5
Validity	8		2		4	14
Veracity		3			8	11
Total Number of Papers per Domain	33	29	27	9	57	155

Figure 2.2: Amount of papers in a domain which use a term at least three times

2.6 Application Domains

In general, application domains are areas where users can create content themselves, which often leads to a fast content growth and is difficult to monitor. The application domains mentioned below are the ones that were the main focus of some papers related to information quality. They are not categorized uniformly, but are taken as they occur in the papers. Application domains which are only mentioned in a short part of a work will be listed as 'others'. Following application domains have been found:

- News
- Reputation and Review Systems
- Healthcare
- eLearning
- Social Media
- Big Data
- Others

2.6.1 News

Papers found in this area can be split into two different types - papers concerning the credibility of news articles and news providers themselves, and papers concerning the credibility of news which is spreading in social media.

1. **News Portals:**
   There have been different approaches to calculate the credibility of single news articles and the credibility of news providers. An interesting approach is presented by [76], whose authors use a sentiment analysis to calculate a credibility value for each article and source. Other approaches use provenance as an indicator [98], or a combination of multiple metrics [107].
2. **Social Media:**
   The major social media platform used for credibility rating has been Twitter, which data are publicly available and which also is a main platform for a quick news spreading. Papers on credibility analysis of

Twitter posts are [7, 24, 63, 66, 73], which use various approaches to reach their goal. One of the studied papers [70] is not using Twitter tweets but is instead ranking blogs by their credibility.

The amount of work on information validation in the area of news stories has increased lately, with a strong focus on news spreading in social media. First works have indeed been about the credibility of news articles and their sources, but as the prime news sources have a steady and often high quality, and thus can be rated manually, the focus has moved to a difficult field of news posted in social media, where the trustworthiness of news posts ranges from very low to very high.

2.6.2 Reputation and Review Systems

There is a vast volume of online service providers in all kinds of areas who offer reputation or review systems related to their service. Some examples are online retail companies like Amazon, which offer product reviews, or discussion forums like Stackoverflow, which offer a reputation system for their users. [69] conducted a large survey of trust and reputation systems for online service provision, including a comprehensive list of problems which tend to occur:

- Low Incentive for Providing Rating
- Bias Toward Positive Rating
- Unfair Ratings
- Change of Identities
- Quality Variations Over Time
- Discrimination
- Change of Identities
- Ballot Box Stuffing

Also available for each item on the list are options to deal with them, but most interesting are the approaches for unfair ratings, as they include a validation of each rating by using credibility scores to users. Most of the work studied by [69] had been released before the year of 2004, which shows that this domain is quite old. A more recent work is described in [25], which

calculates the credibility of product reviews by calculating the quality of its content using various metrics.

2.6.3 Healthcare

In healthcare, two different application domains have been found:

1. **Information quality in healthcare information systems** (HIS) is interesting on an enterprise level, and thus only little work can be found publicly available. Some papers on information quality in general mention healthcare and HIS as a possible application area, but only very few directly relate to it. One is [116], which draws attention to the increasing dependence of healthcare records on information systems and the lack of data quality tools which comprise these new technologies.
2. **Information credibility of online health information** has received much more focus. A reason for this is provided in [68], which analyzes content credibility problems in the area of healthcare. Many people look up health information online, and some of them are not able to recognize unreliable sources. Along with it comes the explosion of content sources and personal health data via websites and forums. There has been theoretical work about trust, with [125] evaluating the use of the word trust in e-health by reviewing related papers. [90] writes about the differences in credibility judgments of online health information in different age groups, [121] analyzes the trust in peer-to-peer healthcare and [26] examines how senior citizens assess the credibility of online health information. Two papers try to automatically assess credibility, with [104] introducing a probabilistic graphical model that jointly learns user trustworthiness of an online health community and the credibility of their medical statements while [126] only focuses on the trustworthiness assessment of single medical statements.

It can be seen that healthcare is a very important area, as the need for health information and the sources credibility is very critical. Most work is from 2010 and later, which is an indicator that the health community has reacted

rather slowly to the shift of the information sources from physicians and books to websites on the internet.

2.6.4 Encyclopedias

Many different encyclopedias exist, most of them very domain-specific, but none of them reaches the scale of the biggest one, Wikipedia. This makes Wikipedia the most used data set for encyclopedias. But regardless of their size, they have one thing in common - their articles are most of the times user generated. Although many encyclopedias already have systems in place to avoid low-quality or intentionally wrong articles, this is still one of the weaknesses of such a user-driven platform. That is why encyclopedias are an important area for information validation. [124], for example, attempts to find a definition for information quality in social information systems, while many others like [4, 93, 95, 119] try to calculate the validity of Wikipedia articles or the trustworthiness of their authors.

2.6.5 eLearning

eLearning is a rarely mentioned area, but still important, as emphasized in [6], which says that "the growing number of available e-learning systems and the commercialisation of these systems highlight the necessity of quality evaluations of online published learning materials." Similar to encyclopedias, there is a need for automatic credibility assessment of user generated eLearning content.

2.6.6 Social Media

Social media are characterized by their user made content and fast growth, both making the media a target for information validation. Though the natural information on social media often tends to spread very fast, regardless whether or not it is correct.

2.6.7 Big Data

Big data is focused from the fact finding community because of its simple nature, which is that it often contains data from many different sources or over a long period of time. One example is [117], which says, as already mentioned in 2.2.11, that veracity is the fourth 'v' of the three 'v' features of big data: volume, velocity and variety.

2.6.8 Others

There are several other application domains, but none of them has been the main focus of a single paper, which is why they are listed here.

Financial area is mentioned as important because of the extreme dependence of financial information on correctness. This has been shown in [89], which created a data set by gathering deep web information of stock data and demonstrated how erroneous it is.

Arbitrary Websites are also a field of interest, as their credibility or trustworthiness can also be rated, as shown by [61].

2.7 Languages

In this section we will analyze which languages are used in the field of information validation research, why a specific language is used and what language differences have to considered.

In Table 2.3, an overview of the used languages can be seen. Entries marked with * are papers which use multiple languages. Noticeable, but hardly surprising is the dominant use of the English language. Firstly since the topic is rather recent, and secondly because it is in the field of computer science, English, as the world language, is of course the first choice for every researcher. In western countries little to no research has been conducted for non-English languages. All papers that actually use different languages for their data sets often use language independent approaches. This is

different in eastern countries like Japan or China, which do a lot of work specific to their own language. Unfortunately, many of them use their native language to write their papers, which leads to them not being included in this thesis.

Language	Papers
English	[70]* [20]* [46]* ... and every other paper not mentioned below.
German	[70]* [20]* [46]*
Japanese	[77] [80] [103] [74] [76] [107]
French	[70]* [46]* [4]*
Spanish	[70]* [46]* [81] [61]
Chinese	[25] [137]
Italian	[70]* [4]* [101]
Portoguese	[46]*

Table 2.3: Languages used in information validation related papers. Entries marked with * are papers which use multiple languages.

The reason why different languages are used often only depends on the country where a work has been written and the availability of the needed data set. Most of the approaches, which will be described in Section 3, are independent of the used language. Where the language matters is the preprocessing of the data (e.g. extraction of facts), which uses NLP techniques and is not a main part of this thesis.

3 Overview of Existing Fact Finding Approaches

In this chapter we will analyze various approaches for assessing the validity of information. The approaches will be classified into two main categories, which are derived from the categories used in the information quality domain [21]: content-based and meta-information-based assessment.

Content-based methods for information validation assessment only use the information itself for the assessment. The approaches can either analyze the information content or compare information with related information. The type of method that can be used depends on the type of information. As the information is in our case often natural language text, these methods can be text analysis methods or methods which derive scores by comparing specific information with other relevant information.

Meta-information-based (also called context-based) methods use the meta-information as an indicator for the validity of information. This meta-information can for example be the date of its creation, a topic, relations to other objects or ratings about the information. Meta-information-based methods are most of the times tailored for a very specific use.

These two methods can be either used solely or together, in which case they will be categorized into meta-information based methods.

The scope of the analyzed methods comprises all text-related information. This includes raw natural language text, semi-structured information like whole websites (in HTML), documents, posts or text snippets with meta-information and completely structured information as in a list of extracted facts. Therefore areas like the validation of images, sound or other non-text-related information will not be further discussed in this thesis.

3.1 Content-based Assessment

In this section we will analyze content-based approaches and data sets used for evaluating those. Content-based information validation is researched in a few domains, which often use different terms to describe a task with basically the same goal. These terms are:

- Fact Finding
- Fact Checking
- Truth Finding
- Truth Discovery
- Data Fusion
- Knowledge Fusion

One basic difference of the approaches listed in the following subsections is in which form they expect their input to be. The majority of approaches already expect the complete information as an input in a structured form. This can be a list of facts about a certain domain, for example books and their corresponding authors. The output of the algorithm will then be which of the facts are wrong and which are right, by just using the given input. Some approaches (3.1.1.1, 3.1.2.3 and 3.1.2.2) do not need a list of facts as an input. Instead, only one single fact must be provided, further data will be retrieved from a data corpus or the web. How these approaches differ from our approach presented in Chapter 4 is elucidated in their corresponding subsection. Although these are two different ways of receiving the needed information, the actual information validation part is the same and thus comparable.

3.1.1 Basic Content-based Approaches

In this subsection, following basic content-based approaches are going to be described:

- Voting
- Counting
- TruthFinder

- Accu, Depen, Accupr,Sim
- CEF-Measure & Copy
- 2-Estimate & 3-Estimate
- Cosine
- Opic
- Solomon
- Trust Propagation
- Investments, Pooled Investments, AvgLog
- Sums & Normalized Sources
- Pop-Accu
- LTM
- LCA
- GTM
- CRH

We consider these approaches to be basic as they only use the facts themselves as input. Any methods which make additional assumptions about the input or which use additional information about the input will be listed under enhanced approaches.

3.1.1.1 Voting

[97] are the first ones to describe an algorithm which exploits redundancy for validating answers in question answering systems. This very simple but effective approach can also be used for general fact validation. This approach will generally be called VOTING and often be the baseline for algorithm evaluation in this area. The basic idea of [97] approach is, "that the number of documents that can be retrieved from the Web in which the question and the answer co-occur can be considered a significant clue of the validity of the answer", which can be generalized with "when trying to find the true fact for a certain object, VOTING chooses the fact that is provided by most websites and resolves ties randomly", as defined by [142]. This approach already utilizes web search engines, but differ from our in Chapter 4 presented approach as it only uses the number of retrieved results instead of the content.

3.1.1.2 Counting

Counting has been used as a baseline by [48]. Counting is similar to Voting but more adaptive as it ignores negative links.

3.1.1.3 TruthFinder

[142] developed an algorithm called 'TruthFinder'. In addition to a simple confidence of facts, the authors of the paper also use trustworthiness for the sources of the facts. So the input of TruthFinder is a bi-partite graph structure, consisting of the source layer and the claim layer. It then calculates the confidence of a claim and the trustworthiness of a source in an iterative process, similar (but not equal) to PageRank [109]. At each iteration, TruthFinder tries to improve its knowledge about their trustworthiness and confidence. It stops when the computation reaches a stable state. The basic calculation TruthFinder is doing is:

- **Trust of a provider:** The average of the confidence of its facts. Additionally, influences between providers are included. An example would be a decease in importance if a source is copying from another source.
- **Confidence of a fact:** When there is only one fact about one object, its confidence is the average of the trust of its providers. When multiple facts about one object exist, they are influencing each other. When two facts are similar, with one having a high confidence, the confidence of the other is increased, too, while it would be decreased if the facts were the opposite.

3.1.1.4 ACCU, DEPEN, ACCUPR, SIM

The approach from [33] is similar to TruthFinder, but uses a different model for calculating the accuracy of the sources. The most important difference is that approaches from [33] consider the dependency between sources. [33] present multiple models, which are all based on the DEPEN approach, which can be seen in Figure 3.1

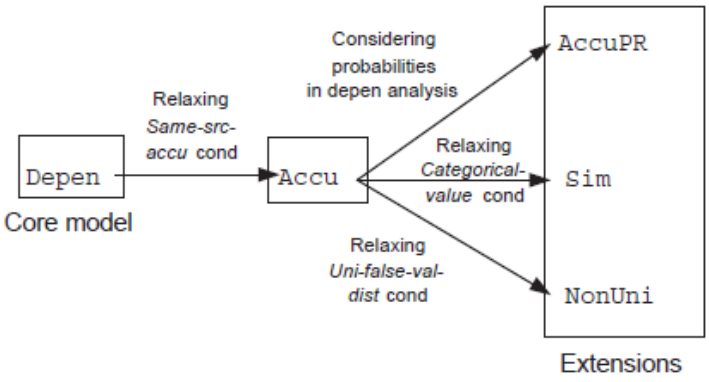

Figure 3.1: Two-layer trust framework and its corresponding three-layer representation by [33]

3.1.1.5 CEF-Measure+Copy

With the CEF approach, [34] attempt to tackle the problem of finding true values and determining the copying relationship between sources, when the update history of the sources is known. The quality of sources is modeled over time by their coverage, exactness and freshness. Based on these measures, a probabilistic analysis is conducted. Firstly, a Hidden Markov Model is used that decides whether a source is a copier of another source and identifies the specific moments at which it copies. Secondly, a Bayesian model is used that aggregates information from the sources to identify the true value for a data item, and the evolution of the true values over time.

3.1.1.6 3-Estimate, 2-Estimate

[48] describe their contribution as follows: "The algorithms estimate the truth values of facts and the trust in sources. They all refine these estimates iteratively until a fixpoint is reached. Their particularities are as follows: 2-Estimates uses two estimators for the truth of facts and the error of views that are proved to be perfect in some statistical sense; 3-Estimates refines

2-Estimates by also estimating how hard each fact is, i.e. the propensity of sources to be wrong on this fact."

3.1.1.7 Cosine

The cosine algorithm from [48] is a heuristic approach for estimating the truth values of facts and the trustworthiness of views, based on the classical cosine similarity measure. Similar to 3-Estimate 3.1.1.6, the estimations are iteratively calculated until a fixpoint is reached. They have also experienced with varying the weight in the calculation (more weight for predictable sources) by using the square instead of the simple absolute value, but with similar results.

3.1.1.8 OPIC

The on-line page importance computation method, short OPIC method, as described by [1], can be used to compute the largest eigenvector and thus can be applied to compute the HITS score. The idea is that each source has two quantities, the cash and the history. When a source is updated, the cash is distributed to its children and the the total quantity is added to the sources history. Its advantage is that it does not matters in which order the sources are updated, or if some sources are updated more often than others, as long they are updated periodically.

3.1.1.9 SOLOMON

The SOLOMON system by [37] contains three components, copying detection, truth discovery and and quality measuring, whereas only the first two are interesting for us. Copying detection is the core of SOLOMON and it proceeds in two steps. The first step, local detection, discovers copying for each pair of sources in isolation of other sources. The second step, global detection, identifies co-copying (multiple sources copying from the same source) and transitive copying (a source copying from a second source, which in turn is copying from a third one), and distinguishes them from

direct copying. Truth discovery is done advancing naive voting in two ways: First, SOLOMON considers the copying relationship and ignores the vote if the provider copies the value from another source. Second, it considers the quality of the sources and gives higher weight to votes from sources of higher accuracy. Based on these two intuitions, a Bayesian analysis is applied and decides on the probability of each observed value being true, considering the one with the highest probability to be the true value.

3.1.1.10 Trust Propagation

Existing approaches use a two-layer architecture that ignores the content and the context in which the source expresses the claim. To overcome this limitation, [126] propose a framework that includes content nodes as an intermediate layer. The framework is a three-tier graph consisting of source, evidence (content), and claim layers, as shown in Figure 3.2 . Each content node represents the evidence given by a source to a claim, and links to one source node and one claim node. This allows the trust framework to explicitly capture the textual context in which a source provides evidence to a claim.

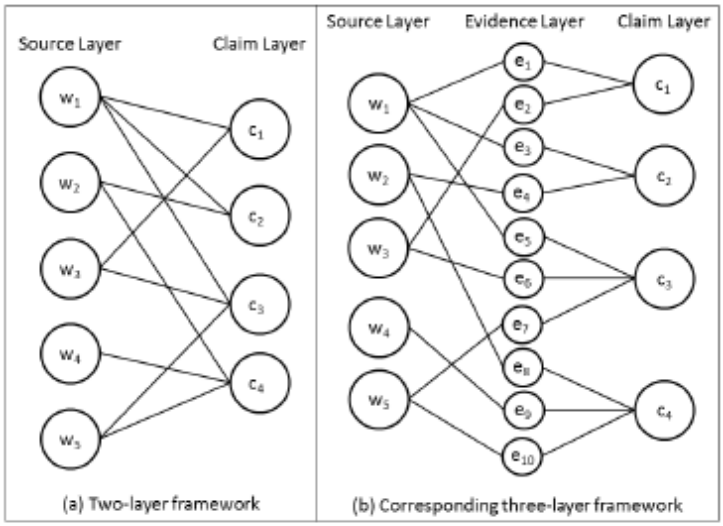

Figure 3.2: Dependency models presented by [33]

3.1.1.11 Investments, Pooled Investments, AvgLog (Prior Knowledge)

[111] introduce a framework for incorporating prior knowledge into any fact-finding algorithm, expressing both general 'common-sense' reasoning and specific facts already known to the user as first-order logic and translating this into a tractable linear program. Additionally, the authors introduce the three new fact-finding algorithms Investments, Pooled Investments and AverageLog.

3.1.1.12 SUMS, Normalized Sources

"SUMS is derived from Hubs and Authorities, where source trustworthiness can be considered the hub score and claim belief the authority score. At each iteration the trustworthiness of each source is calculated as the sum of the belief in its claims, and then the belief score of each claim as the sum of the trustworthiness of the sources asserting it."[112]

Normalized Sources is a variant of SUMS. It additionally normalized the value of a source according to the amount of facts it provides.

3.1.1.13 POP-ACCU (Source Selection)

[35] studied how to select a subset of sources before integration so that we can balance the quality of integrated data and integration cost. It rates the sources with higher accuracy better. For the rating part, the algorithm POP-ACCU based on the ACCU algorithm is presented.

3.1.1.14 LTM

[147] proposed a probabilistic graphical model that can automatically infer true records and source quality without any supervision. In contrast to previous methods, their principled approach leverages a generative process of two types of errors (false positive and false negative) by modeling two different aspects of source quality. In so doing, this is also the first approach designed to merge multi-valued attribute types.

3.1.1.15 LCA

"Latent Credibility Analysis (LCA) constructs strongly principled, probabilistic models where the truth of each claim is a latent variable and the credibility of a source is captured by a set of model parameters." [112]

3.1.1.16 GTM

GTM (Gaussian Truth Model) is a truth-finding method designed specially for handling numerical data. Based on Bayesian probabilistic models, the method can leverage the characteristics of numerical data in a principled way, when modeling the dependencies among source quality, truth, and claimed values [146].

3.1.1.17 CRH

[87] propose to resolve conflicts among multiple sources of heterogeneous data types. The authors model the problem using an optimization framework where truths and source reliability are defined as two sets of unknown variables. The objective is to minimize the overall weighted deviation between the truths and the multi-source observations where each source is weighted by its reliability. The resolve is called Conflict Resolution on Heterogeneous Data (CRH).

3.1.2 Enhanced Content-Based Approaches

In contrast to the basic approaches, the enhanced approaches either make additional assumptions about the input or use additional information. Following enhanced approaches will be described:

- Statistical Fact Checking
- FactChecker
- Community Knowledge
- CATD
- Cluster-based Fact Finder
- Generalizing
- Ensembling
- Attribute Partitioning

3.1.2.1 Statistical Fact Checking

As existing fact-finding models assume availability of structured data or accurate information extraction, which is not always the case, [126] propose a novel, content-based, trust propagation framework that relies on signals from the textual content to ascertain veracity of freetext claims and compute trustworthiness of their sources. The basic steps are: firstly, the extraction of noun-to-noun facts, secondly, the individual fact assessment and thirdly, the document score aggregation. The second part, individual fact assessment, is the part most important in this thesis. [126] accomplish this by generating

four different queries for each fact and extracting facts found in the corresponding top ten results, which are then used to calculate the credibility score for the fact.

3.1.2.2 FactChecker

FactChecker from [50] takes a whole, unstructured document as an input and outputs all found information to facts in the document. For this process, a database which extracted knowledge from DBpedia, YAGO, data.gov and some Twitter feeds is used. At the core of FactMinder lies XR, a data model combining XML and RDF under the single paradigm of annotated documents, and XRQ, its associated query language. No evaluation with a data set has been conducted. This approach is different from the others as it only provides reinforce information for the provided input document, but it does predict of the information is actually right or wrong.

3.1.2.3 Community Knowledge

[127] studied the feasibility of automatically assessing the trustworthiness of a medical claim based on community knowledge and proposed techniques to assign a reliability score for an information nugget based on support over a community-generated collection. The first step is searching for relevant evidence documents (in health forums and mailing lists) that support the claim, to retrieve all occurrences of the treatment relation from a corpus. The second step is scoring individual evidence posts and claims by combining features from retrieved evidence via a scoring functions, and the third step is aggregating the claim scores to compute trustworthiness score for a database of claims. This approach is different to our approach presented in Chapter 4 as it does not use the specific facts found in the used corpus, but instead it uses features of complete text snippets which include the facts.

3.1.2.4 CATD (Confidence-aware truth discovery)

[86] say that existing approaches always overlook the ubiquitous long-tail phenomenon in the tasks, this means that most sources only provide a few claims and only a few sources make plenty of claims. Therefore, the authors propose a confidence-aware truth discovery (CATD) method to automatically detect truths from conflicting data with longtail phenomenon.

3.1.2.5 Cluster-based Fact Finder

[56] derive a model that can evaluate trustworthiness of objects and information providers based on clusters, with the idea that every information provider has its own area of competence (cluster) where it can perform better than others.

3.1.2.6 Generalizing

[113] introduce a generalized fact-finding framework able to incorporate this additional information into the fact-finding process. The key technical idea behind generalized fact-finding is that the relevant background knowledge and contextual detail can be quite elegantly encoded by replacing the bipartite graph of standard fact-finders with a new weighted k-partite graph.

3.1.2.7 Ensembling

[18] compare four ensemble approaches including Simple Bayesian Ensemble, Majority Voting, Uniform Weight and Adjusted Weight ensembles, used together with vairous existing fact-finding methods.

3.1.2.8 Attribute Partitioning

[10] consider the case where there is an inherent structure in the statements made by sources about real world objects, that imply different quality levels of a given source on different groups of object attributes. The authors do not assume this structuring given, but instead find it automatically, by exploring and weighting the partitions of the sets of object attributes, and applying a reference truth finding algorithm on each subset of the optimal partition.

3.1.3 Data sets

In Table 3.1, a summary of existing data sets used to evaluate the content-based approaches is given. This list contains only real world data sets, as the used synthetic data sets only simulate the structure of the real world data sets with an increased amount of entries, and thus are not contributing to any new idea. Instead, they are often just used for performance testing.

Name	Characteristics	Description	Scr
TREC-2001	• Structured • Text • 2,726 entries	492 questions of the TREC-2001 database have been used. For each question, at most three correct and three wrong answers have been randomly selected from the TREC-2001 participants' submissions, resulting in a corpus of 2,726 question-answer pairs.	[97]
AbeBooks	• Structured • Text • 24,364 entries	The data set was extracted by searching computer-science books on AbeBooks.com. In the data set there are 877 bookstores, 1,263 books, and 24,364 listings. Each listing contains a list of authors on a book provided by a bookstore. The correct authors have been chosen using the authors on the cover of the book.	[142]
Hubdub	• Structured • Text • 830 entries	Hubdub is a Web-based prediction market where users can make predictions on future events by answering multiple choice questions. The data set has been constructed from a snapshot of settled questions from May 2009 tagged by the keyword sport. It consists of 357 questions, where each question has between 1 and 20 different answers.	[48]

Table 3.1 – continued from previous page

Name	Characteristics	Description	Scr
Movie Director	• Structured • Text • 108,873 entries	The data set was extracted from bing videos and consists of 15,073 movie entities, 33526 movie director facts, and 108873 claims from 12 sources. 100 movies were randomly sampled for their true directors to be manually labeled.	[147]
US-UK Spelling	• Structured • Text • Entries n.a.	The British National Corpus, Washington Post and Reuters news articles were examined for words with different spelling in the US and UK, taking the sources' usage of a disputed word as a claim and the spelling from a dictionary as the gold standard.	[111]
Large-Scale Knowledge Base	• Structured • Text • Entries n.a.	12 extractors (TXT, DOM, TBL, ANO) were used to extract facts from the web. The knowledge base has a size of 1.6B unique knowledge triples extracted from over 1B Web page.	[36]
GameShow	• Structured • Text • 221,653 entries	The audience of a TV game show could answer multiple choice questions via android application. 38,196 different sources have given 221,653 answers to 2,169 questions. The ground truth information is provided by the TV game show.	[86]
Weather Forecast	• Structured • Text • Entries n.a.	The data contain heterogeneous types of properties. Specifically, weather forecasting data were collected from three platforms: Wunderground, HAM weather, and World Weather Online. On each of them, the forecasts (high temperature, low temperature and weather condition) of three different days were crawled. To get ground truths, we crawl the true weather information for each day. This was done for 20 US cities over a month.	[87]
Exam	• Structured • Text • 30,628 entries	The Exam data set was obtained by aggregating examination results for students applying to the ParisTech program in 2014. The exam is a multiple-choice questionnaire where each question has 5 possible answers, out of which only one is correct. 247 students answered 124 questions.	[10]

Table 3.1 – continued from previous page

Name	Characteristics	Description	Scr
Flight Data	• Structured • Numeric • 27,469 entries	The Flight data set contains 38 Deep Web sources crawled from Google results for keyword search "flight status". Data from 1,200 flights on 12/8/2011 has been collected (flight number,departing airport code, scheduled/actual departure/arrival time, and departure/arrival gate). In total, they provided 27,469 records.	[89]
Stock Deep-Web	• Structured • Numeric • Entries n.a.	The data set contains 55 sources in the Stock domain. We searched 'stock price quotes' and 'AAPL quotes' on Google and Yahoo, and collected the deep-web sources from the top 200 returned results. Every weekday in July 2011, 1,000 stocks have been searched on each data source. Each object is a particular stock on a particular day.	[89]
Flight Deep-Web	• Structured • Numeric • Entries n.a.	Data of 1,200 flights was collected from 38 sources one hour after the latest scheduled arrival time every day in December 2011. Each object is a particular flight on a particular day. The gold standard is the data provided by the three airline websites on 100 randomly selected flights.	[89]
Movie Runtime	• Structured • Numeric • 17.109 entries	The data set contains 603 movies. For each movie, the runtime was collected using Google. 17.109 useful digests have been found, which contain information from 1,727 websites. On average, each movie has 14.3 different runtimes provided by different websites. The runtime provided by IMDB has been considered as the gold standard.	[142]
City Population	• Structured • Numeric • 44,761 entries	Infoboxes for settlements have been collected from Geobox, Infobox Settlement, Infobox City, etc. to obtain 44,761 populations claims qualified by year, with 4.107 authors total. The gold standard is U.S. census data, which provided 308 nontrivial true facts.	[111]

Table 3.1 – continued from previous page

Name	Characteristics	Description	Scr
Stock Returns	• Structured • Numeric • Entries n.a.	A set of stocks were taken that were on the S & P 500 Index on 01/01/2000 and recorded until 01/02/2012. The stock analysts' buy or sell predictions are used as claims about whether each stock will yield a return higher or lower than the baseline S & P 500 return over the next 60 days.	[112]
Indoor Floorplan	• Structured • Numeric • Entries n.a.	The Indoor Floorplan data set contains the distance estimates from users' smart phones for indoor hallways. The hallways distances (129) was manually measured by measuring tapes and used as gold standard.	[86]
Basic Biographies	• Structured • Numeric • 166,733 entries	Infoboxes have been scanned to find 129,847 claimed birth dates, 34,201 death dates, 10,418 parentchild pairs, and 9,792 spouses. The true birth and death dates were extracted from several online repositories (independent and not derived from Wikipedia), which resulted in a total of 2,685 dates as gold standard.	[111]
Manhatten Restaurants	• Structured • Binary • 42,152 entries	5,269 restaurants from 12 web sources that provide information on restaurants in Manhattan have been chosen . Their data has been crawled 8 times in spring 2009. It has been focused on the existence of restaurants (a binary universe). A restaurant counts as closed when the source marks the restaurant as 'CLOSED', or the source removes the restaurant from its list.	[34]
NewsTrust	• Unstructured • Text • 23,164 entries	News data were collected from a community-driven news review website, NewsTrust, in October 2010. Members can rate the various quality aspects of news stories, which gets combined to an overall score from 1 to 5 for each story. This score is considered to be the gold standard. For each news story, the website, author and genre were collected. In total, the data set consists of 23,164 news stories.	[126]
Health Forums	• Unstructured • Text • Entries n.a.	The data set consist of medical claims given in health forum posts. These were extracted by querying for known disease-treatment pairs. The pairs are the claim, the found relevant posts are the evidence and the forums are the source.	[126]

Table 3.1 – continued from previous page

Name	Characteristics	Description	Scr
Medical Treatment	• Unstructured • Text • Entries n.a.	106 valid and 93 invalid treatments across six diseases have been manually collected from various medical web-portals. Based on this gold-standard set of disease-treatment pairs test sets have been constructed. For constructing the test sets, a specific number of valid treatments was randomly sampled for every disease, combined with invalid treatments for that disease.	[127]
Wikipedia Articles	• Unstructured • Text • Entries n.a.	100 Wikipedia articles with 100 or more edits were crawled. The articles included a roughly equal number of featured articles, disputed articles, and randomly selected articles. Each roll back of edits is considered a new article.	[95]

Table 3.1: Data sets used in existing work to evaluate content-based fact finding methods

3.1.4 Analysis

In total, 22 different data sets have been used to evaluate content-based approaches. With 18 data sets, the majority of these data sets are structured, where the content is almost always text or numeric. These structured data sets can be represented as a graph, where the nodes are sources, objects and facts and the links are between objects and facts and sources and facts. Only 4 of the used data sets are unstructured, which means that they are raw text without extracted facts. Each of these unstructured data sets was exclusively used to evaluate one approach, while most of the structured data sets have been used multiple times for evaluating and comparing various approaches.

Table 3.2 shows the approaches for which the data sets were used. When a data set was used for many different approaches, it is an indicator that these approaches were compared in one single work. The most used data set is the ApeBooks data set, which was created in 2008 by [142] as one of the earliest ones for the fact finding domain. The two most used approaches are VOTE and TruthFinder, as these two were often used as baseline for the

evaluation of later approaches. FactChecker has not been evaluated at all, as already mentioned in 3.1.2.2.

	TREC-2001 database	Abe-Books	Movie Runtime	Manhatten Restaurant	Hubdub	US-UK-Spelling	Wiki Basic Biographic	Wiki City Population	DeepWeb Flight	DeepWeb Stock	Stocks Returns	NewsTrust	HealthForums	MedicalTreatment	Wiki Articles	Movie Directors	Flight Data	Large Scale KnowBase	Indoor Floorplan	Game Show	Weather Forecast	Exam
VOTE	x	x	x	x	x	x	x	x	x	x	x					x	x			x		x
TruthFinder		x	x	x	x	x	x	x	x	x	x					x			x	x		
COUNT					x											x						
ACCU		x							x										x	x		
CEF				x																		
3-EST		x			x	x	x	x	x	x	x								x	x		
Cos		x			x				x	x												
SOLOMON		x																				
AvgLog		x				x	x	x	x	x	x					x						
Invest		x				x	x	x	x	x	x					x			x	x		
PooledIn		x				x	x	x	x	x	x					x						
SUMS		x				x	x	x		x	x											
Bayes		x														x						
POP-ACCU		x							x	x							x					
LTM		x					x	x								x						
LCA		x					x	x			x											
GTM							x	x		x									x	x	x	
CRH							x	x	x	x									x	x		
ComKnow														x								
StatFactCheck															x							
CATD							x	x		x									x	x		
Cluster		x					x															
Generalize		x						x														
KnowFusion																		x				
TrustProp												x	x									
Ensembling		x																				
AttPar									x													x
FactCheck																						

Table 3.2: Overview of which data set has been used for evaluation of each fact finding method

3.2 Meta-Information-based Assessment

In this chapter we will analyze meta-information-based approaches and data sets used for evaluating meta-information-based approaches. The examined papers for meta-information-based approaches can be divided into four main data sources:

- Twitter
- Wikipedia
- Forum posts
- Others

As the approaches are tailored to the given data source, they are not easily comparable. It can be seen in the list above that the main data sources are all areas where users have the option to generate the content by themselves. These areas have expanded greatly in the last decade, which made them to a main focus for information validation research. In total, 15 different approaches have been studied.

3.2.1 Meta-Information-based Approaches

3.2.1.1 Twitter, Information Credibility

[24] studied how the credibility of tweets on Twitter could be automatically assessed based on features extracted from them. This work was released in 2011, which made this approach the first published one. The main used features are listed above. This is only to give an overview of which types of features were used for Twitter posts and will not be done for the other approaches, as this would exceed this chapter's scope.

Main features:

- Message-based features:
  - the length of a message
  - whether or not the text contains exclamation or question marks
  - the number of positive/negative sentiment words in a message
  - if the tweet contains a hashtag

  - if the message is a re-tweet
- User-based features:
  - registration age
  - number of followers
  - number of followees ('friends' on Twitter)
  - the number of tweets the user has authored in the past
- Topic-based features:
  - are aggregates computed from the previous two feature sets;
  - the fraction of tweets that contain URLs,
  - the fraction of tweets with hashtags
  - the fraction of sentiment positive and negative in a set.
- Propagation-based features:
  - consider characteristics related to the propagation tree that can be built from the retweets of a message.
  - depth of the re-tweet tree
  - the number of initial tweets of a topic

3.2.1.2 Twitter, Topic-Specific Credibility

[73] also did automatic credibility assessment of tweets on Twitter based on features extracted from them. Half of the features are taken from [24]. Additionally, the credibility assessment was done individually for each topic.

3.2.1.3 Twitter, Alethiometer

[66] introduce Alethiometer, a framework for assessing truthfulness in social media that can be used by professional and general news users alike. This framework is based on text-based analysis capabilities, around three axes: contributor, content and context. For each axis, various parameters are given, which individual normalized scores are added to an aggregated score.

3.2.1.4 Twitter, LDA Features

[63] propose new methods to automatically assess tweet credibility by using two features, 'tweet topic' and 'user topic'. This is done by using the LDA model, which is a well-known generative model for clustering words and documents into mixtures of topics. Because one document (twitter tweet) corresponds to one user, the topic of the document equals the topic of the user.

3.2.1.5 Twitter, Truth in Collective Opinions

[85] describe their work with "It focuses on examining how to reduce the spread of inaccurate information on social media. In particular, we examined the effect of collective opinion on information forwarding in social media environments through an experiment with crowds. In Twitter, an indicator of collective opinion is the number of people who have retweeted a message. The results showed that displaying both retweet counts and collective truthfulness ratings could reduce the spread of inaccurate health-related messages."

3.2.1.6 Blogs, Content Credibility

[70] try to determine the credibility of various blogs on the internet, which have news as their main topic. The credibility of the content of a blog post is measured by first matching it to a specific topic and then measuring the similarity of the blog post to all possible corresponding APA-news articles. The credibility of the author is measured via the credibility of his blogs.

3.2.1.7 Wikipedia, Content-Driven Reputation System

Work on credibility related to Wikipedia has already started in 2007, with [4] content-driven reputation system for Wikipedia authors. In their system, authors gain reputation when the edits they make on Wikipedia articles are preserved by subsequent authors, and they lose reputation when their edits

are rolled back or undone in short order. Author reputation is computed solely on the basis of content evolution, user-to-user comments or ratings are not used.

3.2.1.8 Wikipedia, Credibility via Accessibility

[93] claim that credibility can be measured by the accessibility of a Wikipedia article (accessibility is one of the metrics). They use an UI-accessibility rating framework on Wikipedia articles and the the web pages they link to, to show if Wikipedia has accessibility differences, which would then relate to the credibility of Wikipedia articles. Unfortunately, no evaluation was done if the results correlate with credibility in any way.

3.2.1.9 Wikipedia, Trustworthiness via Edit History

[119] use the Wikipedia edit sequences to determine a community-based user and document trust. The trustworthiness of users and articles are calculated using a machine learning approach on the data set and features from the user, the article, and the edit sequence.

3.2.1.10 Forum Post Credibility 1

[134] use state-of-the-art classification techniques with five feature classes: Surface, Lexical, Syntactic, Forum specific and Similarity features. In detail, they use a support vector machine, namely C-SVM with a Gaussian RBF kernel as implemented by LibSVM in the YALE toolkit.

3.2.1.11 Forum Post Credibility 2

[128] generate a set of features derived from the posting content and the threaded discussion structure for each posting. Similar but not equal to [134], the authors group these features into five categories: relevance, originality, forum-specific features, surface features, and posting-component features.

Using a non-linear SVM classifier, the value of each posting is categorized into one of three levels High, Medium, or Low.

3.2.1.12 Health Forum Post Credibility

[104] assess the Credibility of user-generated medical statements and the trustworthiness of their authors by exploiting linguistic cues and distant supervision from expert sources. Their features are again similar to the approaches described approaches above, but are tailored to the health specific forum.

3.2.1.13 News Articles,Credibility from Sentiment Map

[76] try to provide the credibility of news articles by calculating a sentiment value for news articles. The trends of websites are extracted as average sentiments of the news articles that were written a some topic on each website. The sentiments of news articles are represented by four values calculated in four sentiment scales: 'Bright - Dark', 'Acceptance - Rejection', 'Relaxation - Strain', and 'Anger - Fear'.

3.2.1.14 DeFacto - Deep Fact Validation

Defacto takes an RDF-triple (resource description format) as an input and outputs a confidence score for this triple as well as possible evidence for the fact. The evidence consists of a set of webpages, textual excerpts from those pages and meta-information on the pages. [83]

3.2.1.15 Credibility by Linguistic Indicators

[23] only use linguistic features that can be extracted from written text. The features can be categorized into the eleven groups: quantity, complexity, diversity, specificity, uncertainty, verbal non-immediacy, personalization,

affect, activation, informality and cognitive process. For this work, a scenario with 186 persons was set up where some of them had to lie about a given event. Each person was interviewed about the event, and the the transcribed text got analyzed to find out features which could identify the lying persons.

3.2.2 Data sets

In Table 3.3, a summary of the data sets used to evaluate the meta-information-based approaches is given. In total, 15 different data sets have been used to evaluate meta-information-based approaches. In contrast to the content-based-approaches, each each data set is only used once for the evaluation of a single approach. Each data set is very specific, and the approaches are build around the characteristics of the data sets. Thus, the data sets and their approaches are difficult to compare.

Name	Characteristics	Description	Scr
Twitter Monitor 2010	• Unstructured • Text & Meta-Data • 900,000 entries	Twitter events detected by Twitter Monitor during a 2-months period have been collected. Twitter Monitor is a monitoring system which detects sharp increases in the frequency of sets of keywords found in messages. All tweets matching the query during a 2-day window centered on the peak of every burst have been collected. Each of these sub-sets of tweets corresponds to a topic. The data set consists of over 2,500 topics. Each topic has up to 10,000 tweets.	[24]
Twitter Crawled Topics 2012	• Unstructured • Text & Meta-Data • 1,217,000 entries	Data was crawled from the Twitter streaming API and stored in a relational database (using a python-based crawler for 8 week with 14 different Twitter authentications). 7 popular topics were chosen (Libya, Facebook, Obama, Japanquake, LondonRiots, Horricane, Egypt). 52,000 to 358,000 Tweets and 4 to 37 million users per per topic have been found.	[73]

Table 3.3 – continued from previous page

Name	Characteristics	Description	Scr
Twitter Crawler 2013	• Unstructured • Text & Meta-Data • Entries: n.a.	The data set consists of 10 million users collected from a crawl that has been executed on Twitter content from July 2013 for a period of three months. No further information available.	[66]
Twitter Trendy Tweets 2014	• Unstructured • Text & Meta-Data • 2,000 entries	The data set was created using Twitter's trends API every five minutes to get trendy words in Japan during April 2014. 10 of these 'trends' have been chosen. 200 tweets with trendy words for each trend have been randomly collected from this collection. The tweets where manually labeled by annotators.	[63]
Constructed Data Set	• Unstructured • Text • 42,000 entries	120 health-related statements were selected with two constrains: first, each statement was identified by health professionals as true, debatable, or false; second, the information carried by each statement was familiar to people. Of 120 statements, 40 were true, 40 were debatable, and 40 were false according to health professionals. These statements were then rated by 350 individuals on amazon mechanical turk.	[85]
Popular blogs	• Semi-structured • Text & Meta-Data • Entries: n.a. • German	Blogs were crawled using a high performance web miner. All important parts of a blog were extracted into a structured form. The extracted features are title, date, author, content, language, tags and permanent link. 40 blogs were manually selected by popularity and actuality, and their relation to news topics.	[70]
Wikipedia, Italian and French	• Unstructured • Text & Meta-Data • Over 6 mio entries	The data set consists of the complete Italian Wikipedia, consisting of 154,621 articles and 714,280 filtered revisions (snapshot from December 11, 2005) and the complete French Wikipedia, consisting of 536,930 articles and 4,837,243 filtered revisions (snapshot from October 14, 2006).	[4]
Wikipedia, English	• Unstructured • Text & Meta-Data • 365 entries	The data set consist of 365 Web pages in total, 100 of which come from Wikipedia, whereas 265 are related to external sources. Thus, on average, each Wikipedia article references 2.65 external Web pages.	[93]

Table 3.3 – continued from previous page

Name	Characteristics	Description	Scr
Wikipedia, Simple English	• Unstructured • Text & Meta-Data • Entries: n.a.	Archives of Simple English Wikipedia are available from WikiMedia in an XML format and contain the complete edit histories for most pages. Each article revision contains a unique revision identifier, the editing user, a timestamp, and the full article text at the given snapshot. Each instance contains the ground-truth information of positive reversions as a simple true/false value.	[119]
Nabble Software Forum Posts	• Unstructured • Text & Meta-Data • 1,532 entries	The data set consists of 1,532 rated posts in 1,788 threads from 497 forums, found in the software category of Nabble.com. As users tend to rate extreme (either 1 star or 5 stars), a binary rating is chosen, 947 posts were rated good and 585 bad.	[134]
Slashdot Forum Posts	• Unstructured • Text & Meta-Data • 20,008 entries	The data set is composed of discussion threads from the Slashdot online discussion forum. 200 threads with a maximum of 200 posts each were selected from the 14 subforums on Slashdot. A total of 20,008 rated posts were finally taken from the discussion forum, which were clustered into three groups, namely low, medium, and high, according to their rating value.	[128]
Healthboards Forum Posts	• Unstructured • Text & Meta-Data • 2,800,000 entries	The data were extracted from healthboards.com, an online health community, with 850,000 registered members and over 4.5 million posted messages. 15,000 users and 2.8 million posts were extracted. As ground truth for drug side-effects, data from the Mayo Clinic portal were taken. 2,172 drugs which are categorized into 837 drug families were extracted.	[104]
News Articles, Japanese	• Unstructured • Text • 33,00 entries	The data set consists of about 33,000 news articles from 8 different Japanese online news sites. The articles were collected on September 19, 2007, December 4, 2007, and January 9, 2008.	[76]
News Articles, Trust Propagation	• Unstructured • Text & Meta-Data • Entries: n.a.	See 3.1.1.10. Can be used for single facts or unstructured text.	[126]

Table 3.3 – continued from previous page

Name	Characteristics	Description	Scr
DBpedia, RDF Triplets	• Unstructured • Text & Meta-Data • 600 entries	Facts contained in DBpedia were used as positive examples. For each of the properties considered, positive examples were generated by randomly selecting triples containing the property. In this manner, 600 statements have been collected and verified by checking manually whether it was indeed a true fact. Overall, 473 out of 600 checked triples were facts that could be used as positive examples. The negative examples were derived from positive examples by modifying them while still following domain and range restrictions.	[83]

Table 3.3: Data sets used in existing work to evaluate meta-information-based fact finding methods

3.2.3 Analysis

For the first domain, Twitter, 5 papers regarding credibility assessment have been found. These works are quite new, published from 2011 to 2015. The first four of them [24, 63, 66, 73] are comparable, as they all use feature-based methods to evaluate the credibility. Although they have the same goal and one of them even uses features of another, they all use their own extracted data sets.

In the encyclopedia domain, 3 papers about Wikipedia have been studied. 2 of them [4, 119] use features around Wikipedia articles and their edit history to receive a trustworthiness for authors. Although they have similar methods and similar goals, they also use their own data sets. The third one [93] wants to evaluate if the accessibility (and thus credibility) of Wikipedia articles is different from external websites, but lacks of an appropriate evaluation.

In the third domain, forum posts, again 3 papers have been studied. Each paper has its own data set from a different type of forum, but the methods used are again similar, namely features which partially overlap and a support vector machine for classification.

The last four papers from the domain 'other' are all quite different. [70] evaluates the credibility of German blogs which are related to news topics, [76] creates a sentiment map of Japanese news articles and [23] tries to find linguistic features from written text which can detect if a person is lying or saying the truth. Defacto [83] is similar to content-based approaches, as it tries to validate facts, but it is designed for RDF-triples as input and thus has a lot of additional meta-information to work with.

Each approach has used its own data set. Although for some of the very specific approaches this may makes sense, but for many of them, especially in a single domain like Twitter, the approach could have also be adjusted to be usable for existing data sets. Without any comparison it can not be told if an approach really is as good as claimed by the authors.

4 Dynamic Approach for On-the-Fly Data Set Generation

As shown in the previous Chapter 3, there is a wide range of different approaches for the validation of information. The two fields, content-based approaches and meta-information-based approaches cannot be compared directly, and thus will be treated separately by us.

All existing meta-information-based approaches are tailored for a specific data set of a specific domain. The main differences between the approaches originate from the features provided by a data set, as most of them use machine learning approaches to classify the validity of the input. As none of the used data sets has been made publicly available, it is not easily possible to verify or compare the existing approaches. And the first main challenge for developing a new approach would be to create a data set for a new domain, but the approach itself would again be similar to the existing ones, and thus only generate little new knowledge. These reasons have brought us to focus on the second type, content-based approaches.

The most notable trait of content-based approaches is that even the most simple approach, majority voting (as described by [97]), already shows good results. These results get improved by TruthFinder [142], which is the oldest of the iterative approaches. Most of the approaches presented after these two may have also improved the results, but the differences are often really small. Additionally, many approaches focus on a specific sub task, which of course gives them an advantage for this specific purpose, but otherwise does not really improve to the basic approach. This has led us to the conclusion that it is not necessary to further try to improve fact-finding approaches, as the gain would only be marginal.

One difference between research for content-based approaches and meta-information-based approaches is that many data sets are publicly available. Therefore, also many comparisons of the different approaches were made, which has led to a great transparency. A characteristic all the data sets have in common is that although they are real world data sets, they are still often very simple data sets which do not represent good real world use-cases. Often, a great amount of time has been spend gathering, structuring and cleaning the data for a data set, whereby in a real life example one would most probably only have one single piece of information that needs to be validated. This would exclude the use of most of the existing approaches, as they are built around the existing data sets.

In this work, a first step towards a dynamic approach for on-the-fly data set generation will be presented. This will enable existing fact-finding algorithms to be used for the validation of single facts, without a preexisting data set. For this approach, various parts will be connected to work together, such as querying a search engine, the extraction of text from websites, the extraction of facts from texts, the building of a data set in a suitable format and finally the application of existing algorithms on these data. The difference to existing data sets will be that the generated data set will not be manually cleaned in any way, and thus is expected to have much more wrong information. An example for a similar data set is the AbeBooks data set from [142], which also includes books with their titles and authors, equally to one of the data sets that will be generated by us. Here the difference is that for the AbeBooks data set, the author of each book has been taken from a fixed list of online book stores, which are expected to have very little erroneous entries. For our data set, the authors of books will be extracted from texts of arbitrary websites.

In the sections below we will describe the different parts of the program individually. The basic program structure is displayed in 4.1. The two parts that do already exist are the input, represented by one or more facts that need to be validated (although it could be provided in various forms), and the existing fact finding algorithms. This means that the goal of the program between these two parts is to ultimately output data in a form which can be taken as an input by a fact finding algorithm.

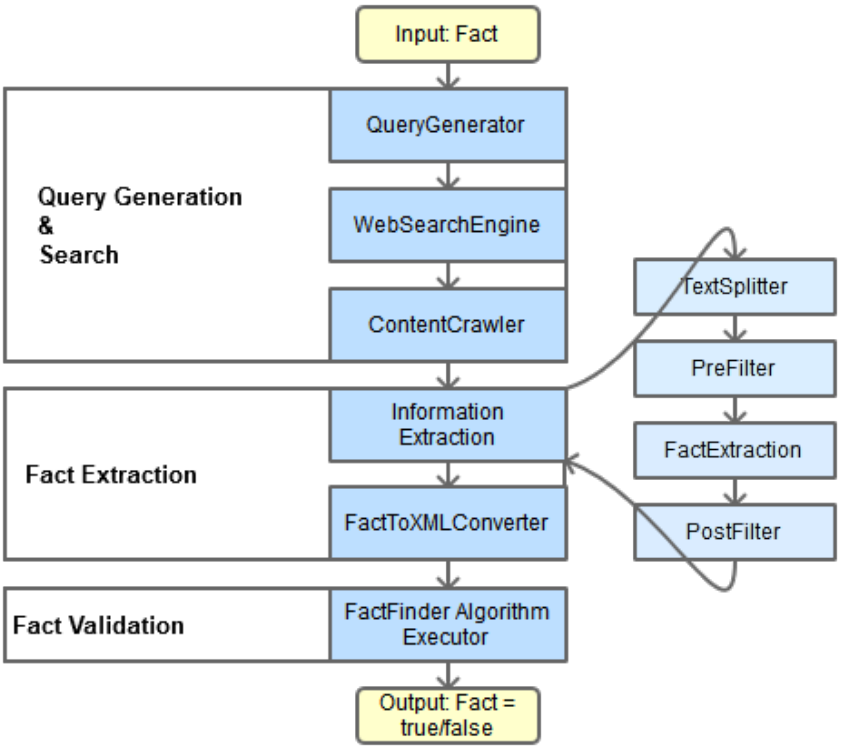

Figure 4.1: Basic program structure for the on-the-fly data set generation

The first step is to define how the perfect input would look like. In general, it consists of multiple sources which provide facts about specific objects, as shown in Figure 4.2.

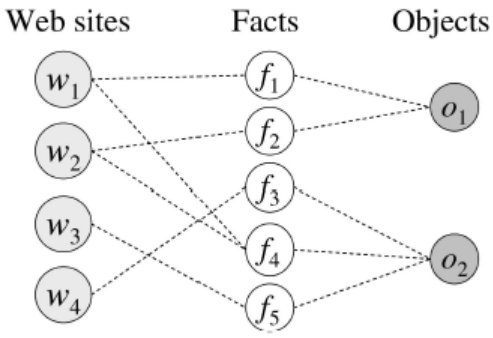

Figure 4.2: Input structure for FactFinder algorithms [142]

The goal of the new data set is to have attributes in a range similar to existing ones. Existing data sets, as listed in Chapter 3.1, have 830 to 221,653 individual facts as entries, provided by 12 to 38,196 different sources, with a big variation of the facts per source ratio. In the movie director data set used by [147], there are 108,873 entries provided by only 12 sources, whereas the movie run time data set from [142] has 17,000 entries provided by 1,727 sources. This means that as a guideline, we want to have about a few hundred entries and a preferably high number of entries per source.

In our case, each source is the domain name of one website. Each entry is a fact extracted from this website, related to the given input query.

As the current defined input is only one fact, this would lead to the data set only containing one single object. To understand the problem see the following example for the input fact 'Microsoft founded by Steve Jobs' that we want to validate:

- **Object 1:** Microsoft founded by
- **Claim 1:** Bill Gates ⇒ Object 1
- **Claim 2:** Steve Jobs ⇒ Object 1
- **Claim 3:** Gates and Allen ⇒ Object 1
- **Website A:** ⇒ Claim 1
- **Website B:** ⇒ Claim 2
- **Website C:** ⇒ Claim 3

With only one object as an input that needs to be verified, each fact would connect to this single object only, and each website would only connect to one single fact. This means that the websites providing facts do not overlap, which would not be a sufficient input for a fact finding algorithm.

To solve this problem, the approach needs to be altered. Instead of using only 1 single fact as an input, multiple similar ones have to be given. This would increase the chance of a website providing facts about multiple objects significantly.

As an object consists of an entity (e.g. 'Microsoft') and a relation (e.g. 'founded by'), the simplest way to achieve this is to vary one of them. So the first way is to vary the relation to find all possible claims about a given entity. This would mean we do not only have the fact 'Microsoft founded by Steve Jobs' as an input, but additionally multiple other facts about the object 'Microsoft', which will help to determine the validity of the first fact. These other relations could be generated automatically, but in a first step they must be provided as additional input by the user. Example relations for 'Microsoft' would be 'founded by', 'headquarter location', 'foundation date', 'foundation location', 'has product' or 'has subsidiaries in'. The second way is to vary the entity (e.g. other tech-companies). Doing both will result in a set of websites providing facts about many different objects.

This means that, for example, the input fact 'Microsoft' 'founded by' 'Bill Gates' will be additionally backed up by a combination of similar objects and relations as shown in the example 4.3 below.

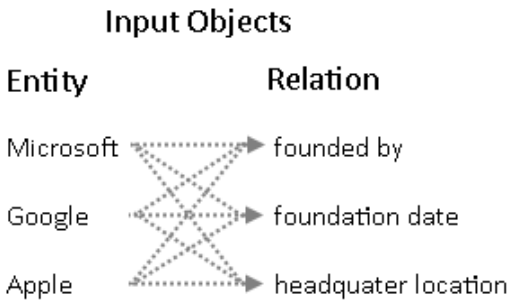

Figure 4.3: Input structure for the presented dynamic approach

4.1 Query Generation & Search

With the limitations defined in the previous section, the program can start to work. Its first part is, as shown in Figure 4.1, the query generation and search step, which consists of the three parts query generation, web search engine and content crawler. These parts will be described in the following subsections. Before defining the query generation step, it is necessary to evaluate which search engines can be used and what boundaries exist for them, thus this step will be described at first, followed by the other two.

4.1.1 Web Search Engine

As already mentioned above, a web search engine will be used for querying the web. The goal is to gather a set of websites most closely related to the original fact. There are several requirements for the search engine, therefore, multiple ones will be tested. For the sake of simplicity, only the most used search engines in the US and Europe will be considered. According to [9], the most used search engine in April 2016 is Google, with a global market share of 71.35%, followed by Bing with 12.37%, Baidu with 7.34% and Yahoo with 7.2%. All others together only reach a market share of 1.74%. As Baidu is a Chinese search engine, and this thesis focuses on the English language, the three web search engines that will be tested are Google, Bing and Yahoo.

The requirements which will be considered are:

1. **Automated queries:** The search engine must allow free access other than through their provided user interface. This means it must provide an API for automated queries.
2. **Query Number:** It must allow enough free automated queries per month so that it does not limit the testing of the developed application. This number should at least be higher than 1,000 queries per month.
3. **Result Number:** The number of received results should be customizable and be at least 50.
4. **Language:** It must be possible to search for results restricted to one single language.

5. **Query Operators:** It must allow at least the simple logical operators 'and', 'or' and 'not' to customize a query.
6. **Speed:** It must be possible to quickly send multiple requests and receive their results.

Point number one is important because all search engines forbid in their terms of service "to access (or attempt to access) any of the Services through any automated means (including use of scripts or web crawlers)"[54]. This means it is not allowed to send automated queries to the default website of a search engine pretending to be a browser. The reason behind this decision is quite obvious: a program sending a query to a search engine will not click on any advertisement, and thus is not profitable for a search engine provider. Additionally, it would, in contrast to a human, be able to send thousands of queries per second, which would unnecessarily stress the servers. So instead, to meet the ToS guidelines, the API of the search engines will be used. Their features are as follows:

4.1.1.1 APIs

Yahoo only provides paid services to their customers. Their free service, the limited search via the Yahoo BOSS API, was discontinued on June 6, 2015. What remained was their paid BOSS JSON Search API for which they charged $1,8 per 1000 queries [140], but even this was discontinued on March 31, 2016 [139], and replaced by Yahoo Partner Ads, which is unfortunately aimed at commercial target audience and not for private users. This makes Yahoo fail at the previously defined requirement number one, and thus it will not be further used.

Bing has the Bing Search API [19] and provides 5000 queries per month for free. To use the Bing Search API, the only thing that needs to be done is to create a Bing Account and request an application id, which can then be used for sending queries to the API. This satisfies requirement number one and two. Bing queries are highly customizable and take additional parameters, with the most important ones being latitude and longitude, the market and the file type. These parameters strongly alter the received results. The queries can be additionally customized with logical operators between multiple terms, as for example 'and' ,'or' and 'not'. The number

of results is 50 at maximum, which again is sufficient for our needs. So the result is a list of websites, where each entry includes the title, the description and the URL. The list can be received in XML or JSON format.

Google has provided the easy to use Google Web Search API, which was officially deprecated as of November 1, 2010 [53]. This API was then replaced with the Google Custom Search API which was originally built to allow website hosts to perform customized searches on their own websites and other specifically defined websites [52]. Fortunately, a workaround makes it possible to also use the Google Custom Search API for searching the entire web. The API provides 100 search queries per day for free, which is about 3000 per month and enough for our purposes. Google limits their free customers by only providing a maximum of 10 results per query, but fortunately it is possible to define from which index you want to get the results, which makes it possible to receive the first 50 results by querying 5 times with different indices. Google also allows to customize each query with additional parameters, with the most important one being 'host language', which boosts documents written in a specific language. It also allows similar logical operators as Bing, again including the most important ones 'and','or' and 'not'. The results can also be received in the JSON format and include each website's title, description and URL.

4.1.1.2 Chosen Search Engine Solution

The only two freely usable search engine APIs are from Google and Bing. After many different test queries for Google and Bing, it became obvious that their returned results are quite different. For each tested query, from the first 50 results websites only 10 to 20 results were the same, which would lead to an overall increase of the result set of at about 80%. As it is desirable to increase the amount of possible sources, the results of both search engines will be merged and all the unique websites will be used in the further steps. Ultimately, this means that we are able to retrieve up to 100 result websites per search query, depending on how many of them are unique. For search engines, the limitations of each search engine to a maximum of 50 results returned per query is not even the bottleneck. Because with all given search

parameters, like a specified language, region and multiple search terms in one query, the amount of matching results often lies in this narrow range.

4.1.2 Query Generation

The goal of the query generation step is to generate web search engine queries and try to get as many possibly relevant websites to the input fact as possible. As already mentioned in the introduction of Chapter 4, we do not have only one input fact which needs to be validated, but also at least one other fact as additional data provider for the algorithms. Depending on how many results are retrieved for the basic query, the amount of additional queries varies.

4.1.2.1 Original Input

First, the query generation for the original input which needs to be validated will be discussed. The input is a fact in the form of:

[Entity] [Relation] [Entity]

An example would be that the fact [cows][eat][grass] should be validated. The first, most basic query is, therefore, 'cows AND eat AND grass', which should return websites that include all three terms. The ulterior motive of this query is that if a fact is wrong, it will yield only a few results.

1st query: Entity AND Relation AND Entity

The goal of the second query is to get the correct fact as a result without using it entirely as input. This means that only the first entity and the relation will be used. If the input fact is correct, this query will return similar results as the first one, if it is false it should instead also return the correct one.

2nd query: Entity AND Relation

The third query has to compensate for the up to now bias towards results including the input facts which emerged with the first query. This means although a fact is wrong and rarely mentioned on the web, the first query

will still return some results. So a query which specifically excludes the second entity should return few results if it is actually correct, and many results if it is wrong.

3rd query: Entity AND Relation NOT Entity

4.1.2.2 Additional Input

Second, the query generation for the additional inputs that help the validation of the original input will be discussed. These inputs have the following form:

[Entity] [Relation]

For the additional inputs, only the basic query 'entity AND relation' will be generated. This ensures that the result set is limited to the most important websites containing the needed data.

Additional queries: Entity AND Relation

All generated queries will be sent to a web search engine, which has already been discussed. The step following after retrieving the web search results is the text extraction step.

4.1.3 Content Crawler

After the previous steps, what we have is a list of a few hundred URLs of websites related to the input fact. What needs to be done next is to download their content and extract the text from the websites. So for this step, the Java HTML Parser 'jsoup' [58], an open source project distributed under the liberal MIT license, will be used, as it comprises all needed tools.

Firstly, the HTML file needs to be retrieved. This is achieved via the jsoup method *connect(String url)*, which creates a new connection, and afterwards *get()*, which fetches and parses the HTML file into a document. Afterwards, this document can be searched through with DOM-like methods, which will be used to extract the text from it.

For the text extraction, it is very important that the document is as clean as possible, since little text chunks from advertisement or any other grammatically wrong pieces of text like lists of words will corrupt the fact extraction step, which is extremely dependent on the correct grammar of its input. Therefore, instead of taking all text from the HTML file, which would include image captions or button labels in an unstructured form, we go through each content node individually, and decide how its content will be used. Ultimately, we distinguish between headlines, paragraphs, lists and tables, and put their content together into one readable text without HTML formatting. This way, the quality of the extracted text increases greatly, which also improves the quality of the following fact extraction.

4.2 Fact Extraction

The second main part of the program (Figure 4.1) is the fact extraction, which consists of the two parts information extraction and the fact to xml converter.

4.2.1 Information Extraction

The fact extraction is the most important but also the most complicated step. As shown in Figure 4.1, it consists of 4 parts which will be described individually: the sentence splitter, a pre-filter, the actual fact extraction with a fact extraction tool, and the post filter.

4.2.1.1 Sentence Splitter

The input we get into the fact extraction step is the text extracted from a website. As most of the information extraction tools are designed to work with short text snippets or single sentences, it is necessary to split the text into sentences. It is important to know that this is not always flawlessly possible. There might be special cases where it seems to the program that a sentence should be split although it should not, or there might be some

human made errors in a sentence which make it unclear to the program what to do. In this case we have to decide if we would rather want to have incorrectly split sentences or incorrectly merged sentences.

Typical problematic sentences include terms like shortcuts, numbers, dates or titles:

"This paragraph includes dates like Feb.20, 2016 , words like U.S and numbers like 1.1, but should not be split. But here, Dr. John.Ok?"

Correctly split the sentences would look like this:

- *"This paragraph includes dates like Feb.20, 2016 , words like U.S and numbers like 1.1, but should not be split."*
- *"But here, Dr. John."*
- *"Ok?"*

A possible approach to split sentences is to use regular expressions combined with the 'String.split' command in Java.

1) *\n—\\.(?!\\d)—(?<!\\d)\\.*

This regular expression will split paragraphs at all new lines and punctual characters, only excluding dots with numbers at both sides. It is very simple, but its disadvantage is that it will split the above paragraph wrongly at 'Feb.20', 'U.S' and 'Dr. John'. The advantage is that mistakes like the missing space between two sentences will be no reason to not split them.

2)*[^.!?\d\ds][^.!?]*(?:[.!?](?![’\d”]?\d\ds—£)[^.!?]*)*[.!?]?[’\d”]?(?=\d\ds—£)*

This regular expression is a more sophisticated one and will detect special cases like 'U.S' or 'Feb.20' and will not split them. But it will also wrongly split 'Dr. John', and overlook to split 'John.Ok'?

A different approach would be to use the java language class BreakIterator, which implements methods for finding the location of boundaries in a text.

3)*BreakIterator iterator = BreakIterator.getSentenceInstance(Locale.US)*:

This approach returns very similar results as the second one, but it takes longer for execution.

4)*DocumentPreprocessor dp = new DocumentPreprocessor(inputText)*:

Fortunately, instead of using these simple approaches, we can also use the Stanford NLP library [99] for this task. Its trained DocumentPreprocessor Class is able to split texts into sentences with a very high precision, being capable of correctly splitting difficult texts.

This makes it very easy to opt for solution number 4, which works as correctly as we need it to.

4.2.1.2 Pre-Filtering

Before starting the information extraction from the text, there is still one additional step to be done, i.e. filtering out useless sentences. As the amount of sentences in a text that are not related to the actual search query is rather big, this would greatly increase the execution time for the information extraction, without giving any disadvantage. Therefore, the ability to filter out sentences that do not match a certain filter has been implemented. In general, only sentences which include the input entities will be taken as input for the information extraction tools. As the HTML to text conversion and the sentence splitter tool already work really clean and well, no further preprocessing is needed.

4.2.1.3 Information Extraction Tools

The information extraction step is expected to be the most critical one. Therefore, the performance of five different tools has been tested. This will help to get a feeling of how different types of extractions look like and which of the tools is most suitable for our task. For the information extraction tools used here, the input is a sentence or a short paragraph and the output consists of structured relation triples. How the fact extraction is effectively done is different for each system, which also leads to different outputs. Below, the differences between the 7 following state-of- the-art open information extraction systems will be described:

- OpenIE: Jan 2016, successor to Ollie (2012)

- ReVerb: Jun 2012
- MitIE: Jan 2015
- ClausIE: Dec 2014
- Stanford-OpenIE: Dec 2015
- CSD-IE: 2013, not publicly available
- LSOE: 2015, not publicly available

The extraction of an example sentence will be shown for each information extraction tool. Keep in mind that this example does not in any way represent the overall performance of the extraction tools. Instead, the purpose of the examples is to show the basic behavior of the tools and the basic structure of their outputs.

- OpenIE: Jan 2016, successor to OllIE (2012)
- ReVerb: Jun 2012
- MitIE: Jan 2015
- ClausIE: Dec 2014
- Stanford-OpenIE: Dec 2015
- CSD-IE: 2013, not publicly available
- LSOE: 2015, not publicly available

The extraction of an example sentence will be shown for each information extraction tool. Keep in mind that this example does not in any way represent the overall performance of the extraction tools. Instead, the purpose of the examples is to show the basic behavior of the tools and the basic structure of their outputs.

KnowItAll OpenIE:

OpenIE 4.0 was authored and developed by people at the University of Washington as part of the KnowItAll project. It contains the principal open information extraction system, which can run over sentences and creates extractions that represent relations in text. OpenIE is the successor to OllIE, which was also developed as part of the KnowItAll project, and consists of the two main components SRLIE [133] and Relnoun [130]. OpenIE creates its extractions from semantic role labeling frames. OpenIE also extends the defintion of open information extractions to include extractions with zero or more arguments [131].

Input	*In 1975, Bill Gates and Paul Allen founded Microsoft, which became the world's largest PC software company.*
Extractions 1	[bill gates and paul allen][founded][microsoft [in 1975]]
Extractions 2	[microsoft][became][the worlds largest pc software company]

Table 4.1: Example extraction of KnowItAll OpenIE

As the example 4.1 suggests, OpenIE handles easily structured inputs very well. The first extraction is a n-ary extraction which perfectly represents the fact. The second extraction is correct, too. OpenIE works very well with many different types of sentences, but it tends to not find temporary arguments in longer sentences, resulting in a very long secondary argument. This means it has a larger priority on being correct than on extracting minimal facts.

KnowItAll ReVerb:

ReVerb is different from the other information extraction tools as it is narrowed down to binary relations only. It automatically identifies and extracts binary relationships from sentences, namely verb-mediated relations. It is designed for web-scale information extraction, where the target relations cannot be specified in advance and speed is important [132][43].

Input	*In 1975, Bill Gates and Paul Allen founded Microsoft, which became the world's largest PC software company.*
Extractions 1	[bill gates] [founded] [microsoft which]
Extractions 2	[microsoft which] [became] [the world s largest pc software company]

Table 4.2: Example extraction of KnowItAll ReVerb

As shown in the example 4.2, ReVerb is more error prone, as it does not distinguish microsoft as own argument. Its correctness increases with shorter

sentences, but as most of the sentences on websites are similar to the example one, ReVerb is not suitable for our approach.

MitIE:

The MitIE tool consists of two parts, a named entity recognition and a binary relation extraction, which basically makes it very different from a standard open information extraction tool. First, the named entity recognition outputs all named entities it can find, and then the binary relation extraction will only use these found named entities to find predefined relations between them. This leads to the found named entities and relations being much less flawed, but also to considerably less named entities and relations being found. Additionally, for each binary relation a trained binary relation model is needed, with 21 of them being available within the MitIE tool. They also provide a tool for training custom models, but this is a lot of work including the need of a good corpus with labeled data [78].

Input	*In 1975, Bill Gates and Paul Allen founded Microsoft, which became the world's largest PC software company.*
Extractions 1	[paul allen] [influenced by] [bill gates]
Extractions 2	[paul allen] [founded organization] [microsoft]

Table 4.3: Example extraction of MitIE

MitIE can only extract predefined relations, which where in this cases people influencing other people and people founding organizations. This is a disadvantage as one is limited to the trained models, but leads to the result of much better extractions.

ClausIE:

The authors describe ClausIE in [29] as a "clause-based approach to open information extraction, which extracts relations and their arguments from natural language text. ClausIE fundamentally differs from previous approaches in that it separates the detection of useful pieces of information expressed in a sentence from their representation in terms of extractions. In more detail, ClausIE exploits linguistic knowledge about the grammar

of the English language to first detect clauses in an input sentence and to subsequently identify the type of each clause according to the grammatical function of its constituents. Based on this information, ClausIE is able to generate high-precision extractions; the representation of these extractions can be flexibly customized to the underlying application. ClausIE is based on dependency parsing and a small set of domain-independent lexica, operates sentence by sentence without any post-processing, and requires no training data (whether labeled or unlabeled)." ClausIE, which is available at [27], makes use of the the Stanford Parser.

Input	*In 1975, Bill Gates and Paul Allen founded Microsoft, which became the world's largest PC software company.*
Extractions 1	[bill gates and paul allen] [founded] [microsoft in 1975]
Extractions 2	[bill gates and paul allen] [founded] [microsoft]
Extractions 3	[microsoft] [became] [the world s largest pc software company]
Extractions 4	[the world] [has] [largest pc software company]

Table 4.4: Example extraction of Stanford ClausIE

As can be seen in the example 4.4, ClausIE works very well and similar to OpenIE, but instead of n-ary extractions it outputs multiple extractions with more minimal content each time, trying to get to the minimal fact. In contrast to the other tools it also extracted the additional fourth fact. ClausIE produces very good results even for larger sentences, with the disadvantage of a long extraction time.

Stanford OpenIE:

Stanford OpenIE extracts structured relation triples from plain text, without knowing the scheme of the relation in advance. The open domain relation triples are extracted by breaking a long sentence into short, coherent clauses, and then finding the maximally simple relation triples [8, 55].

Input	*In 1975, Bill Gates and Paul Allen founded Microsoft, which became the world's largest PC software company.*
Extractions 1	[world] [has] [largest pc software company]
Extractions 2	[bill gates] [founded] [paul allen]

Table 4.5: Example extraction of Stanford OpenIE

The example 4.5 is not representative for Stanford OpenIE, as it does extract correct facts from most of the sentences. In practice it is also very similar to ClausIE.

CSD-IE and LSOE:

The last two state-of-the-art information extraction tools are CSD-IE and and LSOE, but in contrast to the previously mentioned tools they are not publicly available. For completeness, their characteristics will still be shortly described here. LSOE (Lexical-Syntactic pattern-based Open Extractor) is based on generic patterns that identify relations not previously specified [138]. CSD-IE (Contextual Sentence Decomposition) decomposes a sentence into the parts that semantically 'belong together'. By identifying the (implicit or explicit) verb in each such part, facts are obtained [15].

4.2.1.4 Chosen Information Extraction Tool

For the fact finding algorithms to work, the facts in the data set need to be minimal and uniform. Unfortunately, for information extraction tools like ClausIE, this is not the case. A fact can have many different forms, which are not appropriate for further use. For example, the sentence *"A scene from Alfred Hitchcock's film THE MANXMAN"* will result in the extraction [alfred hitchcock][has][film the manxman], which will not be comparable to facts like [the manxman][directed][by alfred hitchcock]. It would be possible to consider such special cases, but given their large volume, this would lead to an incredible amount of work.

We will evade this problem by using an information extraction tool with a uniform output: MitIE. With MitIE, the system will be restricted to the

given pretrained relation models, but for our demonstration purposes this will be sufficient.

4.2.1.5 Post-Filtering

As for standard information extraction tools, after the list of facts has been compiled, it still has many additional facts not related to the input query. In the post-processing step, these are simply filtered out by checking again if any of the keywords from the input query occurs in each of them. But as MitIE only extracts facts for specified relations, the additional post-filtering step has two tasks only; first, to filter out any fact which does not contain any of the input entities, and second, to filter out any duplicate facts of one single source.

4.2.2 Fact Format Converter

At this point, we have a list of fact objects. Each fact object looks like this:

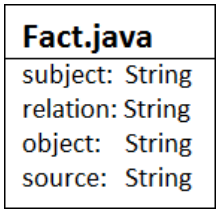

Figure 4.4: Content of the fact object in our implementation

For the fact finding algorithm execution, the project from [48] will be used, which will be further explained in the next section. As an input, data in the following format are needed:

```
<RealData>
    <QueryData>
        <Query id="m0" answer="Subject_Truth_0" name="Object_Relation_0" />
        <Query id="m1" answer="Subject_Truth_1" name="Object_Relation_1" />
          ...
        <Query id="mN" answer="Subject_Truth_N" name="Object_Relation_N" />
    </QueryData>
    <EngineData>
        <Engine id="u0" name="sourceName_0">
            <EngineAnswer answer="Subject_Claim_0" ref="m25" />
            <EngineAnswer answer="Subject_Claim_1" ref="m12" />
            ...
            <EngineAnswer answer="Subject_Claim_n" ref="mX" />
        </Engine>
        ...
        <Engine id="uN" name="sourceName_N">
            <EngineAnswer answer="Subject_Claim_54" ref="m05" />
            <EngineAnswer answer="Subject_Claim_87" ref="m08" />
            ...
            <EngineAnswer answer="Subject_Claim_m" ref="mY" />
        </Engine>
    </EngineData>
</RealData>
```

Figure 4.5: The XML document structure of the output document

It can be seen that instead of having a list of facts, we have a list of unique sources which then provide subject claims to specific object relation pairs. These data would be enough to calculate the probability of a fact being true. Additionally, a list of true facts can be given, which can be used for evaluating the correctness of the algorithm predictions. Therefore, in the fact format conversion step, additional true subjects to the given input object relation pairs will be given, and the data will be converted to the XML format shown in 4.5.

4.3 Fact Validation

The third and last part of the program (Figure 4.1) is the fact validation, in which the generated data set will be used as input for several fact finding algorithms.

4.3.1 Algorithm Execution

Finally, the fact finding algorithms need to be executed on the dynamically created data set. As already mentioned above we used an external tool, 'datacorrob' from [48], to achieve this goal. Their source code is publicly available at [47] and will be adjusted for our needs, which is described in the remainder of this subsection. The project consists of three basic parts:

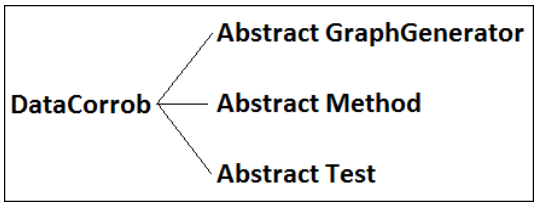

Figure 4.6: DataCorrob project structure

In the first part, 'GraphGenerator', a data set (stored as XML) is taken and converted to an easy to work with graph. There are already some classes defined for existing data sets. In this part, a 'DynamicInputGraphGenerator' class will be added which will deal with our new data set.

In the second part, 'Method', several fact finding algorithms are implemented. They will use the generated graph as an input for their calculation. These algorithms are:

- Voting (3.1.1.1)
- TruthFinder (3.1.1.3)
- OPIC (3.1.1.8)
- NormalizedSources (3.1.1.12)

- CosineSquare (3.1.1.7)
- CosineAbs (3.1.1.7)
- ThreeSteps (3.1.1.6)

In the third part, 'Test', some basic evaluation methods are implemented to show how well an algorithm performs. This part will be extended by the 'SimplePrediction' class, which will just output the predicted correct fact and additionally define if it is truly correct or not.

5 Results and Evaluation

In this chapter first the generation of the new data sets will be described and second the algorithm results for these data sets will be analyzed.

5.1 Data Set Compilation

Next, we will be describe how the actual data sets are acquired, which will then be used to compare different algorithms. To be able to compare the results, four data sets will be acquired, which will range from a simple and small to a bigger and more complex form. As we are limited to relations provided by MitIE, the first thing to do for each data set is to decide for an input fact containing any of these relations.

1) Data Set: Author

As first relation model, '*BOOK_WRITTEN_BY_AUTHOR*' is chosen. The main fact, which we want to have validated, is [Be More Chill][written by][Ned Vizzini]. As already described in Chapter 4, one single fact is insufficient, which is why we use multiple additional backup facts for all data sets. For this data set, 18 additional book titles are provided: 'Promethea', 'The Da Vinci Code', 'Hamlet', 'Invisible Man', 'Twilight new moon', 'Lord of the Flies', 'Barrel Fever', 'Cedar Cove', 'Catching Fire', 'Water for Elephants', 'Life of Pi', 'The Kite Runner', 'Middlesex', 'Atonement', 'American Gods','The Thirteenth Tale', 'Vampire Academy' and 'Cloud Atlas'. All these books were chosen by sticking to famous books which have at least one millions copies sold. So these are in total 19 input facts for the fact validation algorithm. For this input, 21 queries are generated, resulting in 1,851 websites. How these queries are build is described in Chapter 4.1.2. The websites included 4,109 facts matching these relation patterns. The final

data set, cleaned from any facts not related to the given input, consists of 268 facts.

2) Data Set: Founder

As second relation model, '*ORGANIZATION_FOUNDED_BY_FOUNDER*' has been taken. This relation is expected to be a little more difficult, because there are many different people involved with a company who are not the founder. As a main input fact [Walmart][founded by][Karl Wlaschek] is chosen, which is, in contrast to the first data set, a wrong fact. As backup facts, 12 more American companies are used: Samsung, McKesson Corporation, eBay, Goldman Sachs, Berkshire Hathaway, Adidas, Dell, Amazon, FedEx, IBM, News Corporation, and Porsche. As a guideline, only companies today still active in the USA have been chosen. 15 queries led to 1,238 result websites, from which 1930 facts were extracted. After the post processing step, 302 facts remained for the final data set.

3) Data Set: Director

As third relation model, '*FILM_DIRECTED_BY_DIRECTOR*' is chosen. As each movie usually has one director, and the director also stays the same over time, it should be straight forward to distinguish between wrong and right facts. In contrast to the first two data sets, a lot more additional facts are used this time to get a bigger data set. As the main input fact which should be validated is chosen: [The Patriot][directed by][Alfred Hitchcock]. This fact is wrong, and we ultimately want our system to tell us this. As additional backup facts, 42 other movies are used: 'Rear Window', 'The Lambeth Walk', 'The Dark Knight', 'Harry Potter and the Half Blood Prince', 'Pulp Fiction', 'Jumper', 'Life Happens', 'Finding Nemo', 'Inglourious Basterds', 'Gladiator', 'Cloverfield', 'Batman Begins', 'The Bourne Identity', 'Cast Away', 'Sin City', 'Troy', 'Black Hawk Down', 'American Gangster', 'Superbad', 'Memento', 'V for Vendetta', 'The Hangover', 'Kill Bill Vol 1', 'Iron Man', 'Monte Cristo', 'Air America', 'Independence Day', 'Forrest Gump', 'Titanic', 'American Beauty', 'Eyes Wide Shut', 'The Big Lebowski', 'Braveheart', 'American History X', 'The Green Mile', 'Tombstone', 'The Truman Show', 'From Dusk Till Dawn', 'Notting Hill', 'Sixth Sense', 'Air Force One', 'Toy Story', and 'Jumanji'. This in total 43 input facts will lead to 3 queries for the main fact, and one query for each additional fact, resulting in 45 queries. For these queries, a total of 3,824 unique websites were returned,

which is about 89 unique websites per query. The information extraction step led to 7,415 extracted facts, which left 1,091 facts for the data set after the post processing step.

4) Data Set: People

For the last data set, to get an even bigger size without using much more input facts, it is necessary to be able to use more than only one relation. A group of the trained relation models which would fulfill our needs, are relations regarding people:

- *PEOPLE_DEATH_AT_PLACE*
- *PEOPLE_PERSON_ETHNICITY*
- *PEOPLE_PERSON_NATIONALITY*
- *PEOPLE_PERSON_PlACEOFBIRTH*
- *PEOPLE_PERSON_RELIGION*

As this group contains multiple similar relations, if one of them is chosen for the main fact, the others can be used for the 'backup' relations, which will together with additional entities lead to a solid data set. As main input fact which needs to be validated [Marilyn Monroe] [died in] [New York] is chosen, which is again wrong. The used relations are the ones defined above. Additionally to the entitiy 'Marilyn Monroe', 9 other similar entities are used: 'Neil Armstrong', 'Thomas Edison', 'Sigmund Freud', 'Abraham Lincoln', 'Willy Brandt', 'John Locke', 'Arnold Schwarzenegger', 'Wolfgang Mozard' and 'Marie Antoinette'. Combined, this will result in 50 entity relation pairs. This will lead to 3 queries for the main fact, and 49 queries for the additional entity relation pairs, resulting in 52 queries. For these queries, 2,890 unique websites were returned, which is on average 55.6 unique websites per search query. The information extraction step lead to 9,686 facts, of which 1,814 facts ultimately remained for the data set after the post processing step.

These data sets will be used as input for the fact finding algorithms. The results will be discussed in the next section.

Average duration:

The average duration for the generation of a data set depends on the amount of input facts used. The duration for the query generation and search step is

usually really short, as each search engine query is answered in less than a second and the content retrieval is done in parallel with up to 24 concurrent threads, which also leads to only two seconds per 50 query results. The main part of the duration originates from the fact extraction step, which is done sequentially for each website, since it uses a lot of memory. On average, the MitIE tool is able to extract the facts of 10 websites per second. As each query returns a maximum of 100 result websites, the maximum time is 10 seconds for the processing of the results of one query, and for one relational model. For the created data sets, the total creation duration has been between one and seventeen minutes from start to finish.

Truth data:

For evaluation purposes the truly correct facts also need to be provided. This is done manually for each input fact by uniformly taking the information provided by Wikipedia.

5.2 Results

In this section, different features of the tested data sets and the algorithms results will be compared and analyzed by answering certain key questions.

How are the data sets different from existing ones?

In Table 5.1, a short overview of the data sets can be seen. The first two data sets are smaller in size, while the third and fourth ones are a lot bigger.

Data Set	# Input Facts	# Total Facts	# Sources
1) Author	19	268	139
2) Founder	13	302	175
3) Director	43	1091	404
4) People	50	1814	564

Table 5.1: Properties of the dynamically created data sets

It is important to mention the major differences between these data sets and the data sets which have been used in previous works. Firstly, as mentioned above, the main errors, i.e. facts that are wrong, are caused not by sources intentionally providing wrong information, but by the information extraction process itself. Again, this is caused mostly by websites being formatted in a difficult way. A more precise analysis of the types of errors will be provided below. Secondly, the number of sources providing many different facts is rather small. On average, they only provide 2.4 facts per source. The ratio is getting better with an increasing size of the data set. The reason for this is that as a source, only the domain name and not the full website URL is used. The third major difference is the total size of these new data sets, which is at the lower end of the existing data sets. This is caused by the limited amount of website results we are allowed to retrieve per query, and by the limited capabilities of the information extraction tool. But by using multiple input facts, the size of the data sets is still big enough to be comparable.

How well did the algorithms perform?

Table 5.2 shows how many of the input facts were correctly validated for each fact finding method. An input fact is correctly validated when either the predicted fact is the same as the input fact and the input fact is true; or when the predicted fact is different from the input fact, the input fact is false and the prediction is correct. This means that for the evaluation, it does not matter if the provided input fact is true or false, because ultimately it is only checked if the prediction is correct or not. The fields highlighted in green are the methods which had the best results for each data set.

	Correct Outputs			
Method	**1) Author**	**2) Founder**	**3) Director**	**4) People**
Voting	14/19 (74%)	11/13 (85%)	35/43 (81%)	18/48 (36%)
OPIC	14/19 (74%)	10/13 (77%)	36/43 (84%)	17/48 (35%)
NormalizedSources	14/19 (74%)	10/13 (77%)	35/43 (81%)	14/48 (29%)
CosineSquare	10/19 (53%)	10/13 (77%)	37/43 (86%)	20/48 (42%)
CosineAbs	10/19(53%)	10/13 (77%)	37/43 (86%)	20/48 (42%)
ThreeSteps	14/19 (74%)	10/13 (77%)	38/43 (88%)	18/48 (36%)
TruthFinder	15/19 (79%)	10/13 (77%)	37/43 (86%)	19/48 (40%)

Table 5.2: Correctly validated facts for each data set and fact finding method

It can be seen that the number of correctly validated facts is different for each data set, and especially low for the fourth one. That is because in these cases, the facts either simply were not correctly extracted from the websites or many additional wrong facts were extracted because they did also matched the extraction pattern.

Where and why did the algorithms validate facts wrongly?

In Figure 5.1, the prediction results for the author data set can be seen. The value 1 stands for correct predictions, 0 for wrong ones. There were 3 basic types of outputs, which have been marked with a color:

First, marked as red, cases where nearly all algorithms made a wrong prediction. In most of these cases, they all made the same wrong prediction. For the author data set, there are two reasons which cause these wrong predictions. Either the name was extracted only partially or there were only few results, with some of them being wrong.

Second, marked as yellow, cases where only some algorithms have made a wrong prediction. In these cases, some methods have given more weight to specific sources, depending on their way of working. For the books data

set, all methods which made a wrong prediction made the same wrong prediction. The causes for their wrong predictions are the same as above.

And third, marked as white, cases where all methods were right.

	NormalizedSources	TruthFinder	Vote	OPIC	CosineAbs	CosineSquare	ThreeSteps
vampire academy_written by:	1	1	1	1	0	0	1
the kite runner_written by:	1	1	1	1	1	1	1
american gods_written by:	1	1	1	1	0	0	1
be more chill_written by:	0	0	0	0	0	0	1
cedar cove_written by:	1	1	1	1	1	1	1
promethea_written by:	1	1	1	1	1	1	1
lord of the flies_written by:	1	1	1	1	0	0	0
twilight new moon_written by:	1	1	0	0	0	0	0
the da vinci code_written by:	1	1	1	1	1	1	1
cloud atlas_written by:	1	1	1	1	1	1	1
hamlet_written by:	0	1	1	1	1	1	1
invisible man_written by:	0	0	0	0	0	0	0
middlesex_written by:	1	1	1	1	0	0	1
barrel fever_written by:	1	1	1	1	1	1	1
catching fire_written by:	0	0	0	0	0	0	0
water for elephants_written by:	1	1	1	1	1	1	1
atonement_written by:	0	0	0	0	0	0	0
life of pi_written by:	1	1	1	1	1	1	1
the thirteenth tale_written by:	1	1	1	1	1	1	1

Figure 5.1: Correctly predicted facts for the 'Author' data set

In Figure 5.2, the prediction results for the founder data set can be seen. The founder data set is slightly bigger than the previous one, although it has fewer input facts. This leads to a higher fact per input fact ratio. For the founder data set, the reason the two cases where all methods predicted the same wrong fact is simple. In both of them, there have been a lot of extractions, but as many people are involved with a company, a lot of wrong extractions are included. These two cases seemed to be the most confusing ones, since the wrong facts occur much more often than the correct ones.

In the yellow cases, often important people as for example the CEO of a company were mistaken with the founder, but only by some methods.

	NormalizedSources	TruthFinder	Vote	OPIC	CosineAbs	CosineSquare	ThreeSteps
walmart_founded by:	1	1	1	1	1	1	1
amazon_founded by:	1	1	1	1	1	1	1
samsung_founded by:	1	1	1	1	1	1	1
mckesson corporation_founded by:	0	0	0	0	0	0	0
ebay_founded by:	1	1	1	1	1	1	1
goldman sachs_founded by:	1	1	0	1	1	1	1
berkshire hathaway_founded by:	1	1	1	0	0	0	0
adidas_founded by:	0	1	1	1	1	1	1
dell_founded by:	1	1	1	1	1	1	1
fedex_founded by:	1	1	1	1	1	1	1
news corporation_founded by:	1	1	1	1	1	1	1
ibm_founded by:	0	0	0	0	0	0	0
porsche_founded by:	1	1	1	1	1	1	1

Figure 5.2: Correctly predicted facts for the 'Founder' data set

In Figure 5.3, the prediction results for the director data set can be seen. Here, the red marked results had multiple causes, which were:

- The answer is partially true, but not completely (e.g. mr. gibson instead of mel gibson)
- A movie has sequels with a different director (e.g. iron man 1-4)
- There are other movies with a very similar name (e.g. 10 cloverfield lane instead of cloverfield)
- The movie has a remake and with a different director (e.g. gladiator)
- There is a musical with the same name as the movie (e.g. monte cristo)
- Too few results and multiple wrong extractions (e.g. the lambeth walk directed by len lye)

For the director data set, the wrong predictions in the yellow marked cases had a different reason than in the author data set. First, when multiple methods had a wrong result it often was different. The reason for this

is simply that some methods have given more weight to specific sources, depending on their way of working.

	NormalizedSources	TruthFinder	Vote	OPIC	CosineAbs	CosineSquare	ThreeSteps
notting hill_directed by:	1	1	1	1	1	1	1
american beauty_directed by:	1	1	1	1	1	1	1
kill bill vol 1_directed by:	0	1	1	1	1	1	1
v for vendetta_directed by:	1	1	1	1	1	1	1
iron man_directed by:	0	0	0	0	0	0	0
black hawk down_directed by:	1	1	1	1	1	1	1
air america_directed by:	1	1	1	1	1	1	1
the dark knight rises_directed by:	1	1	1	1	1	1	1
memento_directed by:	1	1	1	1	1	1	1
monte cristo_directed by:	1	1	1	1	1	1	1
toy story_directed by:	1	1	1	1	1	1	1
american history x_directed by:	1	1	1	1	1	1	1
titanic_directed by:	1	1	1	1	1	1	1
braveheart_directed by:	0	0	0	1	0	0	1
pulp fiction_directed by:	0	1	0	1	1	1	1
the big lebowski_directed by:	1	1	1	0	1	1	0
harry potter and the half blood prince_directed by:	1	1	1	1	1	1	1
eyes wide shut_directed by:	1	1	1	1	1	1	1
air force one_directed by:	1	1	1	1	1	1	1
sin city_directed by:	1	1	1	1	1	1	1
the patriot_directed by:	1	1	1	1	1	1	1
inglourious basterds_directed by:	1	1	1	1	1	1	1
american gangster_directed by:	1	1	1	1	1	1	1
independence day_directed by:	1	1	1	1	1	1	1
cloverfield_directed by:	1	0	0	0	0	0	0
tombstone_directed by:	1	1	1	1	1	1	1
cast away_directed by:	0	0	0	0	0	0	0
batman begins_directed by:	1	1	1	1	1	1	1
the green mile_directed by:	1	1	1	1	1	1	1
jumanji_directed by:	1	1	1	1	1	1	1
jumper_directed by:	0	1	0	0	1	1	1
sixth sense_directed by:	1	1	1	1	1	1	1
rear window_directed by:	1	1	1	1	1	1	1
the lambeth walk_directed by:	0	0	0	0	0	0	0
gladiator_directed by:	0	0	0	1	0	0	1
finding nemo_directed by:	1	1	1	1	1	1	1
superbad_directed by:	1	1	1	1	1	1	1
troy_directed by:	1	1	1	0	1	1	1
the bourne identity_directed by:	1	1	1	1	1	1	1
from dusk till dawn_directed by:	1	1	1	1	1	1	1
the truman show_directed by:	1	1	1	1	1	1	1
the hangover_directed by:	1	1	1	1	1	1	1
forrest gump_directed by:	1	1	1	1	1	1	1

Figure 5.3: Correctly predicted facts for the 'Director' data set

In Figure 5.4, the prediction results for the people data set can be seen. The people data set is the biggest one, but because of the used relations also the most difficult one. For 48% of all input facts, all methods failed to obtain the correct solution. This high error rate is not caused by bad working methods, but because the correct solution is either not part of the data set at all or does not occur as often as the wrong solutions. There are multiple reasons for this, which all lie in the relations used for this data set. First, for some of the relations, like religion, there could have been more than one during the lifetime of a person, with only the last one being correct. Second, there is also a lot of untrue information spread on the internet, much more than on simple topics such as directors of movies. For example in work about historical people, the true information is often unclear, and therefore different sources claim different things. The third reason for wrong results is that the information extraction tool extracts a large volume of wrong information with these relations. Sentences containing these relations, for example place of birth, are built up similarly to sentences that do not have these relations but still contain locations and people's names. This simply leads to a great amount of wrongly extracted facts. What is more interesting are the yellow marked lines from 5.4, where only some methods had a wrong solution. Here, similar to the previous data set, the wrong results are different from each other, which is caused by the different ways of calculating the trustworthiness of the sources. For example TruthFinder just uses the basic assumption that if a fact is provided by many sources, its trustworthiness is higher, and the trustworthiness of the sources is the average of the confidence of its facts. 3-Estimate additionally uses the trustworthiness of disagreeing sources to reduce the confidence of a fact, which changes the whole process.

	NormalizedSources	TruthFinder	Vote	OPIC	CosineAbs	CosineSquare	ThreeSteps
willy brandt_has place of birth:	0	1	1	1	1	1	0
willy brandt_died at place:	1	1	1	1	1	1	1
neil armstrong_died at place:	1	0	0	0	0	0	1
willy brandt_has nationality:	1	1	1	1	1	1	1
neil armstrong_has place of birth:	1	1	1	1	1	1	1
willy brandt_has ethnicity:	0	0	0	0	0	0	0
arnold schwarzenegger_has nationality:	0	0	0	0	0	0	0
john locke_has nationality:	0	0	0	0	0	0	0
willy brandt_has religion:	0	0	0	0	0	0	0
neil armstrong_has religion:	0	0	0	0	0	0	0
thomas edison_died at place:	0	0	0	0	0	0	0
arnold schwarzenegger_died at place:	0	0	0	0	0	0	0
arnold schwarzenegger_has place of birth:	1	1	1	1	1	1	1
neil armstrong_has nationality:	0	0	0	0	0	0	0
arnold schwarzenegger_has religion:	0	0	0	0	0	0	0
thomas edison_has ethnicity:	0	0	1	0	0	0	0
thomas edison_has place of birth:	1	1	1	1	1	1	1
arnold schwarzenegger_has ethnicity:	0	0	0	0	0	0	0
abraham lincoln_has nationality:	0	1	0	1	1	1	1
thomas edison_has nationality:	0	0	0	0	0	0	0
marilyn monroe_died at place:	0	0	0	1	0	1	1
john locke_died at place:	0	0	0	0	0	0	0
marilyn monroe_has place of birth:	0	0	0	0	0	0	1
marilyn monroe_has religion:	1	1	1	1	1	1	1
marilyn monroe_has ethnicity:	1	1	1	1	1	1	1
wolfgang mozard_died at place:	0	1	0	1	0	1	0
john locke_has ethnicity:	0	0	0	0	0	0	0
neil armstrong_has ethnicity:	0	0	0	0	1	0	0
abraham lincoln_has place of birth:	1	0	0	0	0	0	0
marie antoinette_has nationality:	0	0	0	0	1	0	0
marilyn monroe_has nationality:	0	0	0	0	0	0	0
john locke_has place of birth:	1	1	1	1	1	1	1
wolfgang mozard_has place of birth:	1	1	1	1	1	1	1
abraham lincoln_has religion:	0	0	0	0	0	0	0
thomas edison_has religion:	0	0	0	0	0	0	0
marie antoinette_died at place:	1	1	1	1	1	1	1
john locke_has religion:	0	1	1	1	1	1	0
wolfgang mozard_has nationality:	0	0	0	0	0	0	0
sigmund freud_died at place:	0	0	0	1	0	1	0
marie antoinette_has religion:	0	0	0	0	0	0	0
sigmund freud_has place of birth:	1	1	1	1	1	1	1
marie antoinette_has place of birth:	0	1	1	1	1	1	1
sigmund freud_has nationality:	0	0	0	0	0	0	0
sigmund freud_has religion:	0	1	1	1	1	1	1
marie antoinette_has ethnicity:	0	0	0	0	0	0	0
abraham lincoln_died at place:	0	0	0	0	0	0	0
sigmund freud_has ethnicity:	0	0	0	0	0	0	0
abraham lincoln has ethnicity:	1	1	1	1	1	1	1

Figure 5.4: Correctly predicted facts for the 'People' data set

Do the result differences of the methods coincide with comparisons done in previous works?

Various methods have already been compared to each other on existing data sets, as shown in Table 3.2. In all these comparisons, the results have always been very close, as the simplest method, VOTE, already achieves very high results. This is directly reflected in our new results. TruthFinder only has a little better results than VOTE, while all the other methods also have similar behavior. The only difference that can be found is that for the smaller data sets (author and founder), the results for the COSINE methods are unusual, as they are worse than the baseline. This is most probably caused by the size of these two, which allows bigger variations as it is quite small. In both bigger data sets, they behave properly. The general closeness of all methods' results which can be observed across all four data sets indicates that the more sophisticated methods for fact finding are tailored for specific kinds of data sets, but do not enhance the results for simple ones. The baseline for fact finding algorithms performs really well, and for simple tasks, as for example ours, it is completely sufficient.

What is the reason for the varying results of the different data sets?

On average, the people data set has the lowest success rate, followed by the author, founder and director data set (see Table 5.2). The differences are caused by multiple factors.

The people data set showed the worst results, because the used relations in the input facts are not easy to extract from a text. A person could have lived and traveled through multiple countries, which makes it hard to detect the real nationality. The person's religion might be identified wrongly, and the place of death is not necessarily a well known fact. All these reasons lead to many wrong extractions, which resulted in a high error rate for this data set.

The author data set has an easy input, as usually only one person is mentioned in relation to a book, i.e. the author. Here, the difficulties had from two reasons. Firstly, wrong results were caused by common (i.e. not unique) book titles. Secondly, the entity-relation extraction for books did not work very well, skipping many occurrences of a mentioned fact, which made the author data set the smallest of all created ones.

The founder data set is slightly bigger than the previous one, although it has fewer input facts. This explains the better results, as there are much more extracted facts per given input fact, which makes it easier for the methods to work.

Finally, let us take a look at the movie director data set, which showed the best results. It is similar to the book author data set, but much more input facts were used. Also, the extractor worked very well, which resulted in a big data set. For this data set, the source of error were non-unique titles, due to, for example, the existence of sequels, prequels or remakes of the movie. But these were only minor distractions, and the methods were able to cope with them well.

5.3 Discussion

With the first step towards the dynamic approach we have shown that it is possible to generate a data set on the fly, and to use existing fact finding algorithms to predict the correctness of a given input fact. The used limitations will be analyzed in the next section.

Overall, we have made some very interesting observations:

- The results are on average worse than the results presented in existing evaluations.
- The results of the fact finding methods vary greatly with the type of facts used in the data set.
- The results get worse when the number of facts provided per source is small.
- The baseline algorithms already achieve very good results.

The reason that our results are on average worse than results in existing evaluations simply lies within the data used in the data set. We use more difficult data sets, which is related to point number two, the variation of the method results with the type of the used facts. The results get better when the topic of the input fact is well known, as for example famous movies, as there is more information about it on the web. It also gets better when

there is only one correct, if possible short, answer. For example objects that have long names often tend to be shortened in multiple different ways, which may also be correct but increases the difficulty for the algorithm. Also facts which are not static but instead change over time increase the difficulty extremely. It is easy to determine the date of death of a person, because there is only one which can be true, but it is difficult to determine the religion of a person, as it can change multiple times during a lifetime, and only the last one would be accepted as correct. By the nature of the generated data sets, the number of facts provided per source is smaller than in the existing data sets, which were manually created. This also complicates the work for the algorithms, because this means that there is less proof for a source to be determined trustworthy. The last observation is that the two baseline algorithms achieve very good results, which coincides with existing evaluations.

5.4 Limitations

For the presented approach, following constraints have been made:

- Only free search engine services were used.
- Only free information extraction tools were tested.
- An entity-relation extraction tool was used, limiting the input facts to a given set of predefined relations.
- Multiple additional input facts must be provided by the user to support the main fact.

These limitations are a matter of discussion in future works that are to advance this approach.

Firstly, let us consider the limitation to free search engine services. Here, we are limited to a maximum of 50 results per query by both Bing and Google. To examine how much more results we would be able to retrieve, the standard user interface of Google has been tested manually by inserting some of the previously used queries. The interesting result was that although on the first page, Google often shows that multiple million results were retrieved, the number of real result websites shown is usually only between

50 and 200. This is caused by Google's option to limit the search result to relevant pages only, which removes millions of duplicate and irrelevant websites, which would not contribute to the original query anyway. This means that by using paid services of search engines, it may be possible to increase the results by a factor of 2 to 4, but in no way would the number of results be a thousand websites or more.

Secondly, there is the limitation to the free entity-relation extraction tool. Here, the optimum would be an open information extraction tool which provides minimum facts with as few errors as possible. This would enable the user to use any arbitrary fact as input, and the error-free minimum facts would constitute the perfect data set.

Finally, let us look at the constraint to provide multiple additional input facts. This could be bypassed by only querying for the main input fact and searching all result pages for additional facts they have in common. This is not possible due to the fact that our approach is limited to predefined relations, but it would be if the above described information extraction problem were solved.

6 Conclusion and Future Work

In this thesis, we have analyzed numerous works done in research areas related to information validation, with the goal of providing a big picture of the whole information validation domain. Furthermore, the first step towards a dynamic approach for on-the-fly data set generation has been presented.

6.1 Conclusion

Overall, ten terms that are used to describe the validity of information were found. Although some of them are used more often than others, it was exposed that there is no consistency in the usage of these terms. Single domains tend to stick to similar terms, but not in every case. The most common terms are trustworthiness and reliability in connection to sources of information, correctness in connection to information itself, and credibility in connection to both of them.

In the area of natural language processing, five research areas have been found that deal with the validity of information: information quality, fact finding, question answering, information extraction, and credibility assessment. In these areas, seven big application domains were mentioned, but there are of course also many other smaller ones. These seven are news, reputation and review systems, health care, encyclopedias, eLearning, social media, and big data. The two most important application domains that have been found are social media and news. This comes from the fact that these two domains have a massive amount of publicly available, often user-made content, which tends to be error prone, and is thus ideal for information validation.

Furthermore, it has been found that existing approaches to information validity assessment can be divided into two main groups: content-based methods and meta-information-based methods.

Content-based methods are well researched. The methods have often been compared with each other on many different data sets, where some of them are also publicly available. Newer methods often outperform older ones, but only marginally. In fact, VOTE and TRUTHFINDER, the two approaches often used as baseline, already show very good results for all data sets. To surpass them, data sets with additional information or other special cases are used quite often.

Meta-information-based approaches have a little less research been done about them. In contrast to the previous methods, none of these algorithms or data sets is publicly available, which is also why none of them has been compared to similar methods. This makes it hard to objectively rate the results of these methods.

The presented first step towards a dynamic approach for on-the-fly data set generation, which should make it possible for existing content-based methods to be used to validate single facts, has been implemented with some constraints caused by the functionality of information extraction tools. Firstly, it turned out that facts extracted by open information extraction tools can have too many different forms, which is inappropriate as input for existing fact finding algorithms. So instead, an entity relation extraction tool was used, which limits the possible relations to a predefined set but leads to an output with a fixed form. Secondly, the amount of extracted facts is too small when using only one input fact, as the web search engine queries are limited to a maximum of 50 returned results. This is why additional input facts are used as backup facts, which ensure that the final data set will have a size big enough. Ultimately, four test data sets have been dynamically generated and used as input for existing fact finding methods. The results coincide with previous experiments, and also encourage the conclusion that for content-based methods, the existing baseline methods are good enough for most use cases.

6.2 Future Work

The next step that can be made is to enhance the approach shown in Chapter 4 by working on removing the existing limitations. Firstly, a limitless web search engine API would allow to retrieve a much bigger amount of information related to the given input fact. As this is currently not available for free it has to be newly developed. Secondly, a method for automatically determining multiple facts related to the input fact would greatly improve the usability of the approach, as users would not have to provide them by themselves. This could be done by analyzing the websites containing the input fact and detecting additional facts that also occur on many of the websites. Finally, the entity relation extraction tool can be replaced with an open information extraction tool. But as these tools currently do not provide minimal facts in the quality needed for this approach, they would need a strong improvement. Another topic that is interesting is the meta-information-based assessment of information. Here, no work has been done to compare existing approaches, because the approaches and the existing data sets are unfortunately not publicly available.

Bibliography

[1] Serge Abiteboul, Mihai Preda, and Gregory Cobena. "Adaptive Online Page Importance Computation". In: *Proceedings of the 12th International Conference on World Wide Web*. 2003, pp. 280–290.

[2] Palakorn Achananuparp, Christopher C. Yang, and Xin Chen. "Using Negative Voting to Diversify Answers in Non-factoid Question Answering". In: *Proceedings of the 18th ACM Conference on Information and Knowledge Management*. 2009, pp. 1681–1684.

[3] Gbogboade Ademiluyi, Charlotte E. Rees, and Charlotte E. Sheard. "Evaluating the Reliability and Validity of Three Tools to Assess the Quality of Health Information on the Internet". In: *Patient Education and Counseling* 50.2 (2003), pp. 151–155.

[4] B. Thomas Adler and Luca de Alfaro. "A Content-driven Reputation System for the Wikipedia". In: *Proceedings of the 16th International Conference on World Wide Web*. 2007, pp. 261–270.

[5] Suliman Aladhadh, Xiuzhen Zhang, and Mark Sanderson. "Tweet Author Location Impacts on Tweet Credibility". In: *Proceedings of the 2014 Australasian Document Computing Symposium*. 2014, 73:73–73:76.

[6] Mona Alkhattabi, Daniel Neagu, and Andrea Cullen. "Assessing Information Quality of e-Learning Systems: a Web Mining Approach". In: *Computers in Human Behavior* 27.2 (2011), pp. 862–873.

[7] Majed AlRubaian et al. "A Multistage Credibility Analysis Model for Microblogs". In: *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*. 2015, pp. 1434–1440.

[8] Gabor Angeli, Melvin Johnson Premkumar, and Christopher D Manning. "Leveraging Linguistic Structure For Open Domain Information Extraction". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL*. 2015, pp. 26–31.

[10] Lamine M. Ba et al. "Truth Finding with Attribute Partitioning". In: *Proceedings of the 18th International Workshop on Web and Databases*. 2010, pp. 27–33.

[11] Bartomiej Balcerzak, Wojciech Jaworski, and Adam Wierzbicki. "Application of TextRank Algorithm for Credibility Assessment". In: *Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologies - Volume 01*. 2014, pp. 451–454.

[12] Michele Banko et al. "Open Information Extraction from the Web". In: *Proceedings of the 20th International Joint Conference on Artifical Intelligence*. 2007, pp. 2670–2676.

[13] Daniele Barone, Fabio Stella, and Carlo Batini. "Dependency Discovery in Data Quality". In: vol. 6051. 2010, pp. 53–67.

[14] Hannah Bast, Björn Buchhold, and Elmar Haussmann. "Relevance Scores for Triples from Type-Like Relations". In: *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2015, pp. 243–252.

[15] Hannah Bast and Elmar Haussmann. "Open Information Extraction via Contextual Sentence Decomposition". In: *Proceedings of the 2013 IEEE Seventh International Conference on Semantic Computing*. 2013, pp. 154–159.

[16] Carlo Batini et al. "Methodologies for Data Quality Assessment and Improvement". In: *ACM Computing Surveys* 41.3 (2009), 16:1–16:52.

[17] Behshid Behkamal et al. "A Metrics-driven Approach for Quality Assessment of Linked Open Data". In: *Journal of Theortical and Applied Electronic Commerce Research* 9.2 (2014), pp. 64–79.

[18] Laure Berti-Equille. "Data Veracity Estimation with Ensembling Truth Discovery Methods". In: *2015 IEEE International Conference on Big Data*. 2015, pp. 2628–2636.

[20] Christian Bizer. "Quality-Driven Information Filtering in the Context of Web-Based Information Systems". PhD thesis. Freie Universität Berlin, 2007.

[21] Christian Bizer and Richard Cyganiak. "Quality-driven Information Filtering Using the WIQA Policy Framework". In: *Journal of Web Semantics* 7.1 (2009), pp. 1–10.

[22] Jens Bleiholder and Felix Naumann. "Declarative Data Fusion – Syntax, Semantics, and Implementation". In: *Proceedings of the 9th East European Conference on Advances in Databases and Information Systems*. 2005, pp. 58–73.

[23] Judee K. Burgoon, Lauren Hamel, and Tiantian Qin. "Predicting Veracity from Linguistic Indicators". In: *Intelligence and Security Informatics Conference*. 2012, pp. 323–328.

[24] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. "Information Credibility on Twitter". In: *Proceedings of the 20th International Conference on World Wide Web*. 2011, pp. 675–684.

[25] Chien Chin Chen and You-De Tseng. "Quality Evaluation of Product Reviews Using an Information Quality Framework". In: *Decision Support Systems* 50.4 (2011), pp. 755–768.

[26] Wonchan Choi. "Senior Citizens' Credibility Assessment of Online Health Information: A Proposal of a Mixed Methods Study". In: *Proceedings of the 2012 iConference*. 2012, pp. 620–622.

[28] Chenyun Dai et al. "Assessing the Trustworthiness of Location Data Based on Provenance". In: *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 2009, pp. 276–285.

[29] Luciano Del Corro and Rainer Gemulla. "ClausIE: Clause-based Open Information Extraction". In: *Proceedings of the 22Nd International Conference on World Wide Web*. 2013, pp. 355–366.

[32] Renata Dividino et al. "Querying for Provenance, Trust, Uncertainty and Other Meta Knowledge in RDF". In: *Journal of Web Semantics* 7.3 (2009), pp. 204–219.

[33] Xin Luna Dong, Laure Berti-Equille, and Divesh Srivastava. "Integrating Conflicting Data: The Role of Source Dependence". In: *Proceedings of the VLDB Endowment* 2.1 (2009), pp. 550–561.

[34] Xin Luna Dong, Laure Berti-Equille, and Divesh Srivastava. "Truth Discovery and Copying Detection in a Dynamic World". In: *Proceedings of the VLDB Endowment* 2.1 (2009), pp. 562–573.

[35] Xin Luna Dong, Barna Saha, and Divesh Srivastava. "Less is More: Selecting Sources Wisely for Integration". In: *Proceedings of the VLDB Endowment* 6.2 (2012), pp. 37–48.

[36] Xin Luna Dong et al. "From Data Fusion to Knowledge Fusion". In: *Proceedings of the VLDB Endowment* 7.10 (2014), pp. 881–892.

[37] Xin Luna Dong et al. "SOLOMON: Seeking the Truth via Copying Detection". In: *Proceedings of the VLDB Endowment* 3.1-2 (2010), pp. 1617–1620.

[38] Xin Dong et al. "Knowledge Vault: A Web-scale Approach to Probabilistic Knowledge Fusion". In: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2014, pp. 601–610.

[39] Doug Downey, Oren Etzioni, and Stephen Soderland. "A Probabilistic Model of Redundancy in Information Extraction". In: *Proceedings of the 19th International Joint Conference on Artificial Intelligence*. 2005, pp. 1034–1041.

[40] Oren Etzioni et al. "Open Information Extraction from the Web". In: *Communication of the ACM* 51.12 (2008), pp. 68–74.

[41] Oren Etzioni et al. "Web-scale Information Extraction in Knowitall: (Preliminary Results)". In: *Proceedings of the 13th International Conference on World Wide Web*. 2004, pp. 100–110.

[42] Anthony Fader, Stephen Soderland, and Oren Etzioni. "Identifying Relations for Open Information Extraction". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 2011, pp. 1535–1545.

[43] Anthony Fader, Stephen Soderland, and Oren Etzioni. "Identifying Relations for Open Information Extraction". In: *Proceedings of the Conference of Empirical Methods in Natural Language Processing (EMNLP '11)*. 2011.

[44] David Ferrucci et al. "Building Watson: An Overview of the DeepQA Project". In: *AI magazine* 31.3 (2010), pp. 59–79.

[45] Andrew J. Flanagin and Miriam J. Metzger. "Digital Media and Youth: Unparalleled Opportunity and Unprecedented Responsibility". In: *Digital Media, Youth, and Credibility*. 2008, pp. 5–28.

[46] Giorgos Flouris et al. "Using Provenance for Quality Assessment and Repair in Linked Open Data". In: *Joint Workshop on Knowledge Evolution and Ontology Dynamics*. 2012.

[48] Alban Galland et al. "Corroborating Information from Disagreeing Views". In: *Proceedings of the Third ACM International Conference on Web Search and Data Mining*. 2010, pp. 131–140.

[49] Liang Ge et al. "Multi-source Deep Learning for Information Trustworthiness Estimation". In: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2013, pp. 766–774.

[50] François Goasdoué et al. "Fact Checking and Analyzing the Web". In: *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*. 2013, pp. 997–1000.

[51] David C. Gondek et al. "A Framework for Merging and Ranking of Answers in DeepQA". In: *IBM Journal of Research and Development* 56.3.4 (2012), 14:1–14:12.

[56] Manish Gupta, Yizhou Sun, and Jiawei Han. "Trust Analysis with Clustering". In: *Proceedings of the 20th International Conference Companion on World Wide Web*. 2011, pp. 53–54.

[57] Olaf Hartig. "Trustworthiness of Data on the Web". In: *Proceedings of the STI Berlin & CSW PhD Workshop*. Citeseer. 2008.

[59] Bernd Heinrich and Mathias Klier. "Metric-based Data Quality Assessment: Developing and Evaluating a Probability-Based Currency Metric". In: *Decision Support Systems* 72 (2015), pp. 82–96.

[60] Bernd Heinrich, Mathias Klier, and Marcus Kaiser. "A Procedure to Develop Metrics for Currency and Its Application in CRM". In: *Journal of Data and Information Quality* 1.1 (2009), 5:1–5:28.

[61] Christopher Horn et al. "Using Factual Density to Measure Informativeness of Web Documents". In: *Proceedings of the Nordic Conference of Computational Linguistics* (2013).

[62] Marie Iding, Brent Auernheimer, and Martha E. Crosby. "Towards a Metacognitive Approach to Credibility". In: *Proceedings of the 2Nd ACM Workshop on Information Credibility on the Web*. 2008, pp. 75–80.

[63] Jun Ito et al. "Assessment of Tweet Credibility with LDA Features". In: *Proceedings of the 24th International Conference on World Wide Web*. 2015, pp. 953–958.

[64] Kristo Ivanov. "Quality-Control of Information: On the Concept of Accuracy of Information in Data-Banks and in Management Information Systems". PhD thesis. Stockholm: The Royal Institute of Technology KTH, 1972.

[65] Ian Jacobi, Lalana Kagal, and Ankesh Khandelwal. "Rule-based Trust Assessment on the Semantic Web". In: *Proceedings of the 5th International Conference on Rule-based Reasoning, Programming, and Applications*. 2011, pp. 227–241.

[66] Eva Jaho et al. "Alethiometer: A Framework for Assessing Trustworthiness and Content Validity in Social Media". In: *Proceedings of the 23rd International Conference on World Wide Web*. 2014, pp. 749–752.

[67] Wojciech Jaworski, Emilia Rejmund, and Adam Wierzbicki. "Credibility Microscope: Relating Web Page Credibility Evaluations to Their Textual Content". In: *Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT) - Volume 01*. 2014, pp. 297–302.

[68] Colleen Jones. "Will Content Credibility Problems Flatline Health Innovation?" In: *Interactions* 19.5 (2012), pp. 22–25.

[69] Audun Jøsang, Roslan Ismail, and Colin Boyd. "A Survey of Trust and Reputation Systems for Online Service Provision". In: *Decision Support Systems* 43.2 (2007), pp. 618–644.

[70] Andreas Juffinger, Michael Granitzer, and Elisabeth Lex. "Blog Credibility Ranking by Exploiting Verified Content". In: *Proceedings of the 3rd Workshop on Information Credibility on the Web*. 2009, pp. 51–58.

[71] Joseph M. Juran and Frank M. Gryna. *Juran's Quality Control Handbook*. McGraw-Hill, 1974.

[72] Aditya Kalyanpur et al. "Structured Data and Inference in DeepQA". In: *IBM Journal of Research and Development* 56.3.4 (2012), 10:1–10:14.

[73] Byungkyu Kang, John O'Donovan, and Tobias Höllerer. "Modeling Topic Specific Credibility on Twitter". In: *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces*. 2012, pp. 179–188.

[74] Takuya Kawada et al. "Web Information Analysis for Open-domain Decision Support: System Design and User Evaluation". In: *Proceedings of the 2011 Joint WICOW/AIRWeb Workshop on Web Quality*. 2011, pp. 13–18.

[75] Daisuke Kawahara, Sadao Kurohashi, and Kentaro Inui. "Grasping Major Statements and Their Contradictions Toward Information Credibility Analysis of Web Contents". In: *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*. 2008, pp. 393–397.

[76] Yukiko Kawai et al. "Using a Sentiment Map for Visualizing Credibility of News Sites on the Web". In: *Proceedings of the 2Nd ACM Workshop on Information Credibility on the Web*. 2008, pp. 53–58.

[77] Yutaka Kidawara. "Information Credibility Analysis of Web Content". In: *Proceedings of the 2Nd ACM Workshop on Information Credibility on the Web*. 2008, pp. 3–4.

[79] Shirlee-ann Knight and Janice Burn. "Developing a Framework for Assessing Information Quality on the World Wide Web". In: *Informing Science Journal* 8 (2005), pp. 159–171.

[80] Takuya Kobayashi et al. "Modeling and Analyzing Review Information on the Web Focusing on Credibility". In: *Proceedings of the 2009 ACM Symposium on Applied Computing*. 2009, pp. 1316–1317.

[81] Dimitris Kontokostas et al. "Test-driven Evaluation of Linked Data Quality". In: *Proceedings of the 23rd International Conference on World Wide Web*. 2014, pp. 747–758.

[82] Yang W. Lee et al. "AIMQ: A Methodology for Information Quality Assessment". In: *Information and Management* 40 (2002), pp. 133–146.

[83] Jens Lehmann et al. "DeFacto - Deep Fact Validation". English. In: *The Semantic Web – ISWC 2012*. Vol. 7649. 2012, pp. 312–327.

[84] Yangyong Zhu Li Cai. "The Challenges of Data Quality and Data Quality Assessment in the Big Data Era". In: *Data Science Journal* 14.2 (2015), pp. 1–10.

[85] Huaye Li and Yasuaki Sakamoto. "Computing the Veracity of Information through Crowds: A Method for Reducing the Spread of False Messages on Social Media". In: *System Sciences, 2015 48th Hawaii International Conference on*. 2015, pp. 2003–2012.

[86] Qi Li et al. "A Confidence-aware Approach for Truth Discovery on Long-tail Data". In: *Proceedings of the VLDB Endowment* 8.4 (2014), pp. 425–436.

[87] Qi Li et al. "Resolving Conflicts in Heterogeneous Data by Truth Discovery and Source Reliability Estimation". In: *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*. 2014, pp. 1187–1198.

[88] Xian Li et al. "Scaling up Copy Detection". In: *Data Engineering , 2015 IEEE 31st International Conference on*. 2015, pp. 89–100.

[89] Xian Li et al. "Truth Finding on the Deep Web: Is the Problem Solved?" In: *Proceedings of the VLDB Endowment* 6.2 (2012), pp. 97–108.

[90] Q. Vera Liao and Wai-Tat Fu. "Age Differences in Credibility Judgments of Online Health Information". In: *ACM Transactions on Computer-Human Interaction* 21.1 (2014), 2:1–2:23.

[91] Hui Lin et al. "A Context Aware Reputation Mechanism for Enhancing Big Data Veracity in Mobile Cloud Computing". In: *2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing*. 2015, pp. 2049–2054.

[92] Jimmy Lin and Boris Katz. "Question Answering from the Web Using Knowledge Annotation and Knowledge Mining Techniques". In: *Proceedings of the Twelfth International Conference on Information and Knowledge Management*. 2003, pp. 116–123.

[93] Rui Lopes and Luis Carriço. "On the Credibility of Wikipedia: An Accessibility Perspective". In: *Proceedings of the 2Nd ACM Workshop on Information Credibility on the Web*. 2008, pp. 27–34.

[94] Marianela G. Lozano et al. "Towards Automatic Veracity Assessment of Open Source Information". In: *2015 IEEE International Congress on Big Data*. 2015, pp. 199–206.

[95] Amr Magdy and Nayer Wanas. "Web-based Statistical Fact Checking of Textual Documents". In: *Proceedings of the 2Nd International Workshop on Search and Mining User-generated Contents*. 2010, pp. 103–110.

[96] Bernardo Magnini et al. "Comparing Statistical and Content-Based Techniques for Answer Validation on the Web". In: *In Proceedings of the VIII Convegno AI*IA*. 2002.

[97] Bernardo Magnini et al. "Is It the Right Answer?: Exploiting Web Redundancy for Answer Validation". In: *Proceedings of the 40^{th} Annual Meeting on Association for Computational Linguistics*. 2002, pp. 425–432.

[98] Erik Mannens et al. "Automated Trust Estimation in Developing Open News Stories: Combining Memento & Provenance." In: *COMPSAC Workshops*. 2012, pp. 122–127.

[99] Christopher D. Manning et al. "The Stanford CoreNLP Natural Language Processing Toolkit". In: *Association for Computational Linguistics (ACL) System Demonstrations*. 2014, pp. 55–60.

[100] Mausam et al. "Open Language Learning for Information Extraction". In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. 2012, pp. 523–534.

[101] Mario Mezzanzanica et al. "A Model-based Evaluation of Data Quality Activities in KDD". In: *Information Processing & Management* 51.2 (2015), pp. 144–166.

[102] Jerzy Michnik and Mei-Chen Lo. "The Assessment of the Information Quality with the Aid of Multiple Criteria Analysis". In: *European Journal of Operational Research* 195.3 (2009), pp. 850–856.

[103] Hisashi Miyamori. "Assisting the Validity Assessment of Items Based on Composition Similarity". In: *Proceedings of the ACM Multimedia 2009 Workshop on Multimedia for Cooking and Eating Activities*. 2009, pp. 15–22.

[104] Subhabrata Mukherjee, Gerhard Weikum, and Cristian Danescu-Niculescu-Mizil. "People on Drugs: Credibility of User Statements in Health Communities". In: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2014, pp. 65–74.

[105] J.William Murdock et al. "Textual Evidence Gathering and Analysis". In: *IBM Journal of Research and Development* 56.3.4 (2012), 8:1–8:14.

[106] J.William Murdock et al. "Typing Candidate Answers using Type Coercion". In: *IBM Journal of Research and Development* 56.3.4 (2012), 7:1–7:13.

[107] Ryosuke Nagura et al. "A Method of Rating the Credibility of News Documents on the Web". In: *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2006, pp. 683–684.

[108] Felix Naumann et al. "Data Fusion in Three Steps: Resolving Inconsistencies at Schema-, Tuple-, and Value-level". In: *In Bulletin of the Technical Commitee on Data Engineering*. 2006, pp. 21–31.

[109] Lawrence Page et al. *The PageRank Citation Ranking: Bringing Order to the Web.* Technical Report 1999-66. Stanford InfoLab, 1999.

[110] Jeff Pasternack and Dan Roth. "Generalized Fact-finding". In: *Proceedings of the 20th International Conference Companion on World Wide Web*. 2011, pp. 99–100.

[111] Jeff Pasternack and Dan Roth. "Knowing What to Believe (when You Already Know Something)". In: *Proceedings of the 23rd International Conference on Computational Linguistics*. 2010, pp. 877–885.

[112] Jeff Pasternack and Dan Roth. "Latent Credibility Analysis". In: *Proceedings of the 22Nd International Conference on World Wide Web*. 2013, pp. 1009–1020.

[113] Jeff Pasternack and Dan Roth. "Making Better Informed Trust Decisions with Generalized Fact-finding". In: *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*. Vol. 3. 2011, pp. 2324–2329.

[114] Veronika Peralta. *Data Freshness and Data Accuracy: A State of the Art*. 2006.

[115] Leo L. Pipino, Yang W. Lee, and Richard Y. Wang. "Data Quality Assessment". In: *Communications of the ACM* 45.4 (2002), pp. 211–218.

[116] Sabitha Rajan and Srini Ramaswamy. "On the Need for a Holistic Approach to Information Quality in Healthcare and Medicine". In: *Proceedings of the 48th Annual Southeast Regional Conference*. 2010, 39:1–39:5.

[117] Barna Saha and Divesh Srivastava. "Data quality: The other Face of Big Data". In: *2014 IEEE 30th International Conference on Data Engineering*. 2014, pp. 1294–1297.

[118] Markus Schaal et al. "Information Quality Dimensions for the Social Web". In: *Proceedings of the International Conference on Management of Emergent Digital EcoSystems*. 2012, pp. 53–58.

[119] Jeffrey Segall et al. "Assessing Trustworthiness in Collaborative Environments". In: *Proceedings of the Eighth Annual Cyber Security and Information Intelligence Research Workshop*. 2013, 52:1–52:4.

[120] Fatimah Sidi et al. "Data quality: A Survey of Data Quality Dimensions". In: *2012 International Conference on Information Retrieval Knowledge Management*. 2012, pp. 300–304.

[121] Elizabeth Sillence, Claire Hardy, and Pam Briggs. "Why Don'T We Trust Health Websites That Help Us Help Each Other?: An Analysis of Online Peer-to-peer Healthcare". In: *Proceedings of the 5th Annual ACM Web Science Conference*. 2013, pp. 396–404.

[122] Besiki Stvilia et al. "Research Project Tasks, Data, and Perceptions of Data Quality in a Condensed Matter Physics Community". In: *Journal of the Association for Information Science and Technology* 66.2 (2015), pp. 246–263.

[123] Katsumi Tanaka. "Web Search and Information Credibility Analysis: Bridging the Gap Between Web1.0 and Web2.0". In: *Proceedings of the 3rd International Conference on Ubiquitous Information Management and Communication*. 2009, pp. 39–44.

[124] Roman Tilly et al. "What is Quality of Data and Information in Social Information Systems? Towards a Definition and Ontology". In: *Thirty Sixth International Conference on Information Systems* (2015), pp. 1–21.

[125] Laurian Vega, Enid Montague, and Tom DeHart. "Trust in Health Websites: A Review of an Emerging Field". In: *Proceedings of the 1st ACM International Health Informatics Symposium*. 2010, pp. 700–709.

[126] V.G. Vinod Vydiswaran, ChengXiang Zhai, and Dan Roth. "Content-driven Trust Propagation Framework". In: *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2011, pp. 974–982.

[127] V.G. Vinod Vydiswaran, ChengXiang Zhai, and Dan Roth. "Gauging the Internet Doctor: Ranking Medical Claims Based on Community Knowledge". In: *Proceedings of the 2011 Workshop on Data Mining for Medicine and Healthcare*. 2011, pp. 42–51.

[128] Nayer Wanas et al. "Automatic Scoring of Online Discussion Posts". In: *Proceedings of the 2Nd ACM Workshop on Information Credibility on the Web*. 2008, pp. 19–26.

[129] Xianzhi Wang et al. "Approximate Truth Discovery via Problem Scale Reduction". In: *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. 2015, pp. 503–512.

[134] Markus Weimer, Iryna Gurevych, and Max Mühlhäuser. "Automatically Assessing the Post Quality in Online Discussions on Software". In: *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*. 2007, pp. 125–128.

[135] Philip Woodall, Alexander Borek, and Ajith Kumar Parlikad. "Data quality assessment: The Hybrid Approach". In: *Information & Management* 50.7 (2013), pp. 369–382.

[136] Minji Wu and Amélie Marian. "Corroborating Answers from Multiple Web Sources". In: *Tenth International Workshop on the Web and Databases*. 2007.

[137] Youzheng Wu, Xinhui Hu, and Hideki Kashioka. "Mining Redundancy in Candidate-bearing Snippets to Improve Web Question Answering". In: *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*. 2007, pp. 999–1002.

[138] Clarissa Castella Xavier, Vera Lucia Strube de Lima, and Marlo Souza. "Open information extraction based on lexical semantics". In: *Journal of the Brazilian Computer Society* 21.1 (2015), pp. 1–14.

[141] Mohamed Yahya et al. "ReNoun: Fact Extraction for Nominal Attributes". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. 2014, pp. 325–335.

[142] Xiaoxin Yin, Jiawei Han, and P.S. Yu. "Truth Discovery with Multiple Conflicting Information Providers on the Web". In: *IEEE Transactions on Knowledge and Data Engineering* 20.6 (2008), pp. 796–808.

[143] Xiaoxin Yin and Wenzhao Tan. "Semi-supervised Truth Discovery". In: *Proceedings of the 20th International Conference on World Wide Web*. 2011, pp. 217–226.

[144] Amrapali Zaveri et al. "Quality Assessment for Linked Data:A Survey". In: *Semantic Web Journal* 7.1 (2015), pp. 63–93.

[145] Amrapali Zaveri et al. "User-driven Quality Evaluation of DBpedia". In: *Proceedings of the 9th International Conference on Semantic Systems*. 2013, pp. 97–104.

[146] Bo Zhao and Jiawei Han. "A Probabilistic Model for Estimating Real-valued Truth from Conflicting Sources". In: *Proceedings of the 10th International Workshop on Quality in Databases*. 2012.

[147] Bo Zhao et al. "A Bayesian Approach to Discovering Truth from Conflicting Sources for Data Integration". In: *Proceedings of the VLDB Endowment* 5.6 (2012), pp. 550–561.

Internet Sources

[9] Net Applications.com. *Search Engine Market Share 2016*. 2016. URL: https://www.netmarketshare.com/search-engine-market-share.aspx?qprid=4&qpcustomd=0.

[19] Bing. *Bing Search API*. 2016. URL: https://datamarket.azure.com/dataset/bing/searchweb.

[27] Aaron Culich. *Stanford ClausIE*. 2016. URL: https://github.com/aculich/clausie.

[30] LLC Dictionary.com. *Dictionary.com Unabridged*. 2016. URL: http://dictionary.reference.com.

[31] LLC Dictionary.com. *Roget's 21st Century Thesaurus, Third Edition*. 2016. URL: http://www.thesaurus.com.

[47] Alban Galland. *Data Corroboration Project*. 2016. URL: http://gforge.inria.fr/projects/datacorrob/.

[52] Google. *Google Custom Search API*. 2016. URL: https://developers.google.com/custom-search/.

[53] Google. *Google Search API*. 2016. URL: https://developers.google.com/web-search/?csw=1.

[54] Google. *Google Terms of Service*. 2016. URL: http://www.google.com/policies/terms/archive/20070416-20120301/.

[55] Stanford NLP Group. *Stanford OpenIE*. 2016. URL: http://nlp.stanford.edu/software/openie.html.

[58] Jonathan Hedley. *jsoup HTML parser*. 2016. URL: https://jsoup.org/.

[78] Davis E. King. *MITIE*. 2016. URL: https://github.com/mit-nlp/MITIE.

[130] University of Washington's Turing Center. *KnowItAll chunkedextractor*. 2016. URL: https://github.com/knowitall/chunkedextractor.

[131] University of Washington's Turing Center. *KnowItAll OpenIE*. 2016. URL: https://github.com/knowitall/openie.

[132] University of Washington's Turing Center. *KnowItAll ReVerb*. 2016. URL: https://github.com/knowitall/reverb.

[133] University of Washington's Turing Center. *KnowItAll SRLIE*. 2016. URL: https://github.com/knowitall/srlie.

[139] Yahoo. *Yahoo BOSS Search*. 2016. URL: https://developer.yahoo.com/boss/search/.

[140] Yahoo. *Yahoo Policies*. 2016. URL: https://policies.yahoo.com/us/en/yahoo/terms/product-atos/boss/pricing/index.htm.

Appendices

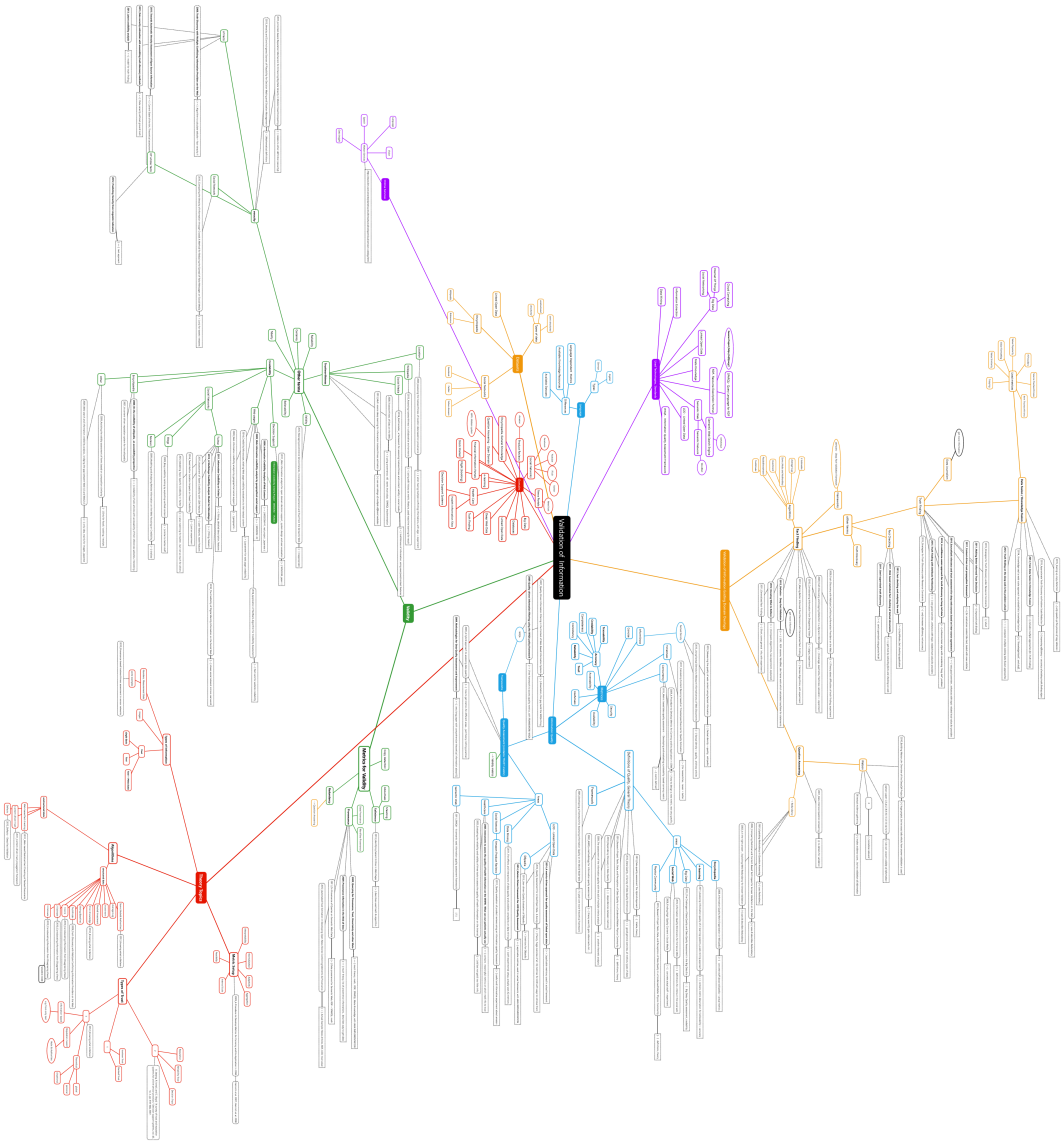

Figure .1: Mind map which links all found papers to specific topics

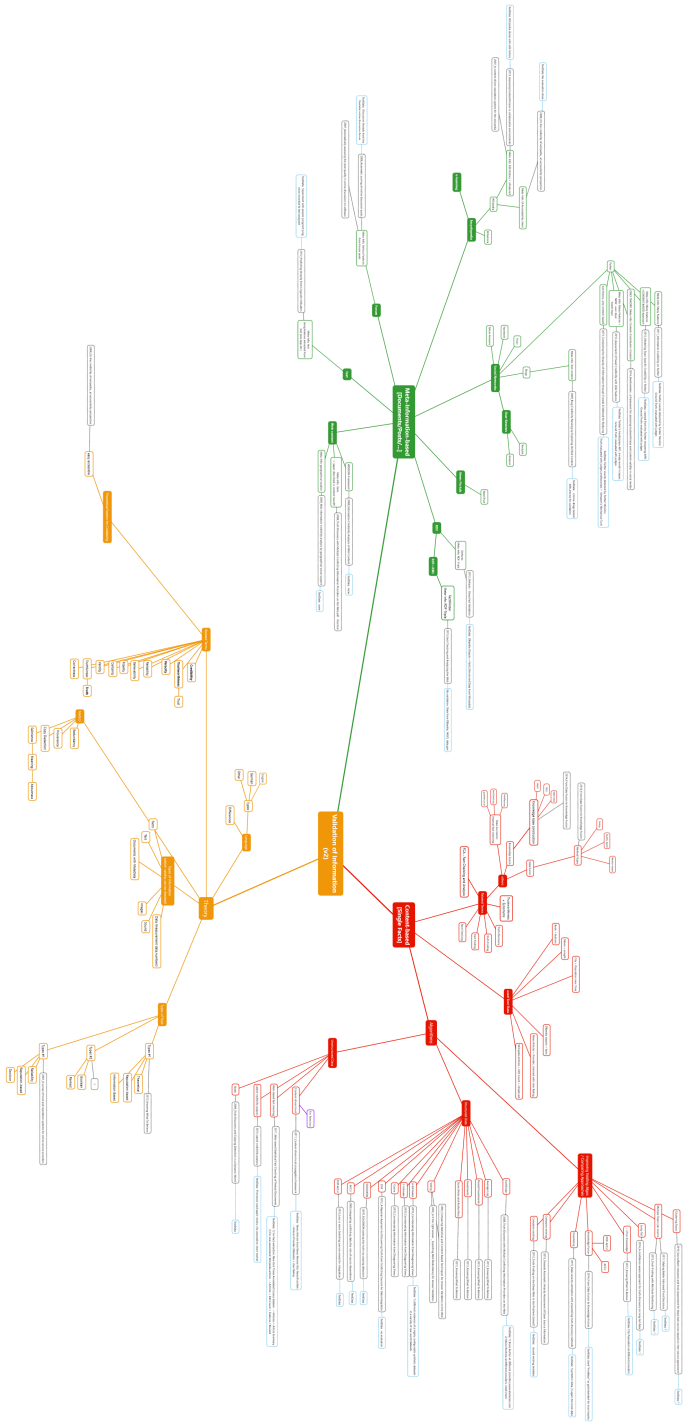

Figure .2: Mind map which links important papers to important topics