Using Knowledge from Wikipedia

to Improve Document Classification

Martin Boroš

European Masters Program

in Language and Communication Technologies

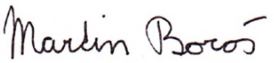
University of Groningen, Shanghai Jiao Tong University

Supervisors: Dr. Gosse Bouma, University of Groningen

Tianfang Yao Ph.D., Shanghai Jiao Tong University

ACKNOWLEDGEMENT

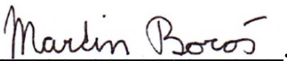
I, Martin Boroš, assure that all the matter of this report is my work unless specifically referred and given proper credit. I would like to extend my appreciation and gratitude to my instructors who have been a sign of inspiration and true guides helping me through the process of conducting this research. I am indebted to The European Masters Program in Language and Communication Technologies for the opportunity it has given me. During the course of the program I had the privilege to meet many people who have in one way or the other inspired me, supported me or provided me with insights and ideas. The research was made possible by the help of a number of people who directly or indirectly encouraged me, and guided me to finish this work at this acceptable level.

Signature: 

Date: 17 August 2012

DECLARATION

I, Martin Boroš, declare that to the best of my knowledge and effort, this report is my work, and does not use the material of others unless proper credit is given. I have taken possible steps to avoid any misuse of information or and other such mishaps. Any issues that may arise are accidental and not intentional in anyway.

Signed:  . Date: 17 August 2012.

Abstract

The intention of the research is to understand how document classification can be improved using Wikipedia. A variety of literatures have been reviewed to get a better understanding of the topic, and to study details of the functions of Wikipedia and the effect it can have on document classification. Spoken and written communication further facilitated our technological advancement. Starting with the early Greeks, the encyclopedia emerged as a modern form of culture and knowledge transmission and has served as an important tool for the collecting and archiving of knowledge allowing future generations the opportunity to build on prior developments rather than continual rediscovery. The popularity of Wikipedia has also grown and currently (as of October, 2011) ranks fifth in overall global web traffic (“Alexa Top 500 Global Sites,” n.d.). Web users looking for information on any topic will likely come across a Wikipedia article fairly quickly. Keeping this in mind we fail to reject H1 = Document classification can be improved using Wikipedia, H3 = Algorithm models are the factors that contribute to document classification, and H6 = the current position document classification is good. We propose methods for automatic enrichment of bag of words representation using knowledge from Wikipedia.

TABLE OF CONTENTS

CHAPTER 01: INTRODUCTION

The intention of the research is to understand how document classification can be improved using Wikipedia. A variety of literatures have been reviewed to get a better understanding of the topic, and to study details of the functions of Wikipedia and the effect it can have on document classification. This chapter will shed light on some basic information so the reader can have a better understanding of some details so the discussion and the conclusion are clearer. Humans seem to have an innate desire and need to transmit knowledge to future generations. Such knowledge transmission clearly had evolutionary benefits as well. Types and forms of stone tools, for example, demonstrate the impact of culture and shared knowledge. Although the emergence of language cannot be exactly determined, it clearly coincided with a long period of technological and cognitive development of early man (Renfrew, Frith & Malafouris, 2008).

Spoken and written communication further facilitated our technological advancement. Starting with the early Greeks, the encyclopedia emerged as a modern form of culture and knowledge transmission and has served as an important tool for the collecting and archiving of knowledge allowing future generations the opportunity to build on prior developments rather than continual rediscovery. The digital age has rapidly accelerated our ability to create record and share knowledge as well as offer new opportunities for collaboratively constructing knowledge. Wikipedia is a unique approach that relies on crowd-sourcing knowledge, but while hugely popular it remains to be seen if this approach can result in a legitimate source of authoritative knowledge or will degenerate into a form of cultural tribalism (Arazy, Nov, Patterson, & Yeo, 2011) over who owns the truth.

Background of the research

The encyclopedia has largely been taken for granted and not greatly studied (Kafker, 1981). Nevertheless, the encyclopedia has come to represent the pinnacle of general knowledge transmission and it has become common for school age children and adults to pick up a volume when looking for information on a topic. Dating back to at least the ancient Greeks, the encyclopedia has gone through a number of changes culminating in the modern, multi-volume, alphabetically organized sets we see today such as the English language Encyclopedia Britannica or The World Book Encyclopedia. Venerable print encyclopedias such as these are now being challenged by digital encyclopedias that rely on the efforts of unnamed volunteers to add, edit and update content. Currently, the most well-know example is Wikipedia which, since its initial release in 2001, has grown to over 3.7 million articles in English and over 20 million articles in over 280 languages.

The popularity of Wikipedia has also grown and currently (as of October, 2011) ranks fifth in overall global web traffic ("Alexa Top 500 Global Sites," n.d.). Web users looking for information on any topic will likely come across a Wikipedia article fairly quickly. However, the open approach to editing content and even creating new articles, a process in which anyone can edit nearly any page (some pages are locked at times for various reasons), has resulted in a steady stream of criticism regarding quality, accuracy, authority of its authors, susceptibility to vandalism, and overall legitimacy as a reliable reference tool. Despite a growing body of research suggesting that Wikipedia content is generally credible (Chesney, 2006) and not significantly more error-prone than print encyclopedias (Arazy et al., 2011; Chesney, 2006; Giles, 2005; Magnus, 2006; Rajagopalan et al., 2010; Rector, 2008; Rosenzweig, 2006), no encyclopedia is ever going to be completely free of errors, but digital encyclopedias have the

potential to respond much more quickly when mistakes are found. Shortly after publication of the Nature study (Giles, 2005) it was reported that all the identified errors were fixed (Snow, 2006).

Conversely, an interesting example of the persistence of outright false information in a print encyclopedia is the story of the so-called Piltdown man, or Dawson's Dawn Man, reportedly found by Charles Dawson between 1908 and 1912. Dawson claimed the skull was an example of a heretofore unknown missing link in human evolution that contained a mix of modern human and primate features. The discovery was widely reported at the time and accounts of what was later proven to be a hoax remained in such venerable resources as the Encyclopedia Britannica until as recently as 1949 – or nearly 40 years after the initial report (Collison, 1966; "Glacial Epoch," 1949; "Sources and authorities for English history," 1949). Interestingly, accounts of the hoax are now included in both Britannica ("Piltdown man," 2002) and Wikipedia ("Piltdown Man," n.d.). In a somewhat ironic passage referring to the Piltdown man, the 1922 version of the Encyclopedia Britannica stated,

Research on text categorization

Research on text categorization is recently moving from plain text documents to semi-structured XML documents. The XML mining track (Decoyer and Gallinari, 2007) of the INEX is pioneering this research using the Wikipedia corpus where the categories are non-overlapping i.e., each document belongs to only one category. One of the prominent approaches that are applied for text categorization is the vector space model. The efficiency of this model relies heavily on the creation of category profiles that consist of features and their weights and the appropriate selection of similarity measure. These profiles are built from the training-set and are

used in identifying the category of the test document. Features that represent the category are carefully chosen by identifying the term's distribution at various levels of the given document and category. The most widely used feature weighting schemes are Term Frequency (TF), Inverse Document Frequency (IDF) and Document Frequency (DF) which assign the weights based on the presence of the terms within the positive category only (Salton and Buckley, 1988).The categories other than the positive category to which the document belongs are called negative categories. The presence of terms in such negative categories affects the weight of features. This factor is considered in defining the Relevance Frequency (RF) measure (Yang et. al., 2002; Lan el.al, 2005).

Once the category profiles are built using the feature selection methods, relevance measure is used to compare the profile of the test document with the category profiles. The unknown test document is classified to the category having the highest degree of relevance. The appropriate selection of the relevance measure has a great impact on the effectiveness of categorization. Though existing feature selection approaches make use of the negative category distribution in creating category profiles, effectively it reduces the importance of features in positive category alone. These positive features are tested for their presence in the unknown test document, whose degree of similarity is to be measured. The main drawback in these measures is that, the closeness of two sets of positive features alone is compared. However, the negative features which represent the deviation between two text objects also needs to be considered in order to come up with a balanced similarity measure. Hence, there is a need for representing both positive and negative features within the test document to be categorized. The Wikipedia XML documents contain several structural elements such as links, sections and their titles, tables and references each having different levels of importance. Each document is an article about a

particular topic. Our approach makes use of the vector space model for representing the features in a category. Positive features are split into two categories viz. pure and shared depending on their contribution to a particular category. To be able to identify how document classification can be improved using Wikipedia file classification, it is essential to see and understand what document classification is in detail.

Document classification

The concept of "classification" is used most often at the same time and in the value of the process and within the meaning of the result, i.e. understood as a group and as a result that is received in the scheme. In order to limit the classification process and its outcome documents we use the two terms:

Classification of documents

Classification of documents is the process of ordering and distribution of documents in classes in order to reflect the relationship between them and the drawing up of the classification scheme.

Classification

Classification is a system of subordination used as a means of establishing links between the classes of documents, as well as orientation in their diversity. The structure of the classification is usually presented as a table or schema.

Importance of classification

It is important to keep in mind that classification is a method of learning. Without it, it is impossible to study the diversity of types of documents, organize them, to establish the differences between the types of documents available on various grounds. A comprehensive classification reflects the pattern of the documents, reveals the links between them, helping them to navigate in any set that serves as a basis for ordering in document systems. It is important for the theory and practice of document communication activities. To carry out the classification of documents is especially important to keep in mind, as a minimum, the following provisions:

1. sign by which produce division, called the base of the division, and formed in this concept - the members of the division;
2. the same division should be on the same basis;
3. the sum of all the members of the division should be equal to the total amount of the dividend concept, i.e. division must be exhaustive, nor failure, nor redundant divisions are not allowed (the requirements of proportionality);
4. members of the division should mutually to exclude each other's (the requirements mutually exclusive);
5. Members of the division should be closest to the concept of dividend shall not be allowed to jump from the next in a series of division in the distant or above the underlying (the requirement of continuity).

Process of classification

There are three approaches to the problem of text classification. First, the classification is not always carried out using a computer. For example, in the ordinary library books are assigned

subject headings manually by a librarian. Such a manual classification does not apply in cases where it is necessary to classify a large number of documents at high speed. Another approach is to write the rules by which the text can be attributed to one or another category. For example, one such rule might look like this: "if the text contains the words of the derivative and the equation, then take it to the category of mathematics." The specialist, who is familiar with the subject area and have the skills to write a regular expression, can make a set of rules, which are then automatically applied to incoming documents for their classification. This approach is better than the last, because the classification process is automated, and therefore the number of documents processed is practically unlimited. Moreover, the construction of the rules by hand can give better classification accuracy than the machine learning (see below). However, the creation and maintenance of the rules up to date (for example, to classify the news with the name of the current president of the country, the appropriate rule should be changed from time to time) requires a constant effort specialist.

Finally, the third approach is based on machine learning. In this approach, a set of rules or, more generally, the criterion for deciding a text classifier is calculated automatically from the training data (in other words, training the classifier). Training data - this is some good samples of documents from each class. In machine learning remains the need for manual partitioning (the term means the process of assigning markup document class). But the markup is much simpler task than writing rules. In addition, the markup can be made in the ordinary mode of using the system. For example, e-mail program may be able to mark messages as spam, thus forming a training set for the classifier - filter unwanted messages. Thus, the classification of texts based on machine learning, is an example of supervised learning where the teacher acts as a person, given a set of classes and mapping out a training set.

Like any job classification, categorization of documents can be achieved by way supervised or unsupervised. Supervised mode, pre-defined elements will be used to classify a document: it may be an index or a dictionary of words corresponding to a particular class and used to pre-labeled documents. Unsupervised mode, it is on the learning phase that will build the training of the classifier, and its subsequent performance. The process used in document classification systems of numerical algorithms. The most successful are those based on SVM or of Boosting (based on AdaBoost). Other methods of similarity measure (such as cosine similarity) or probabilistic (the naïve Bayesian classifiers) can also be implemented. This is used in the most efficient systems – a combination of several classification systems by counting up a voting method.

Research Aim

The aim of the research is to understand how document classification can be improved using knowledge from Wikipedia.

Research Objectives

The objectives of the research are mentioned below.

1. To identify how document classification can be improved using Wikipedia.
2. To identify the factors that contribute to document classification.
3. The current position document classification.

Research Questions

The research will aim at trying to answer the following questions.

1. How can document classification can be improved using Wikipedia.
2. What are the factors that contribute to document classification.
3. What is the current position document classification.

Hypothesis

H_1 = Document classification can be improved using Wikipedia.

H_2 = Document classification cannot be improved using Wikipedia.

H_3 = Algorithm models are the factors that contribute to document classification.

H_4 = Algorithm models are the factors that do not contribute to document classification.

H_5 = the current position document classification is poor.

H_6 = the current position document classification is good.

CHAPTER 02: LITERATURE REVIEW

A number of literature was reviewed to understand and identify different factors in which Wikipedia can influence document classification. The topics in Wikipedia documents are discussed in different sections. Terms that are distributed along different structural elements across different documents tend to contribute more to a category than terms which are limited to a specific structural element alone. We define Structural Term Frequency (STF) for a particular document as a combination of Term Frequency (TF) and Structural Frequency (SF). Term Frequency calculates the number of times the term occurs in a document irrespective of the nature of the distribution within structural elements. The presence of features in large number of sections and titles within a document implies the higher contributing power. Hence, TF is incremented by SF which is the number of sections and titles the term appears.

Mean Free Path Based Categorization

Motivated by the Collision Theory, the research defines the effective collision score by considering three types of collisions in the test document viz. Free-Free transition, Free-Bound transition and Bound-Bound transition as shown in equation (5). For each category, we have identified the features that are shared among other negative categories. Features whose weights are higher in the positive category compared to any of the negative categories and features that are unique to a particular category combine to form the Pure Feature Set (PFS). The rest of the positive features are grouped in Shared Feature Set (SFS). The set of terms that are not in the positive and shared feature set, but are unique to the respective negative categories are grouped as Negative Feature Set (NFS).

Free-Free transition results in the energy gain for the particular electron under consideration, as it absorbs the photon with which it interacts. We have mapped this with successive positive feature transitions. When two positive features either from the PFS or SFS form a chain, it reflects higher relevance towards the positive category. This fact is reflected in equation (5) by giving higher weight to Free-Free transition chains. However, Bound-Bound chains do not contribute heavily as the features in the NFS are not confined to a particular negative category alone. As a result of a collision, the feature having the higher weight gains more weight according to the weight of the nearby term. We have used exponential growth function for applying the feature gain for each transition. Free-Free transitions are calculated until a negative feature is encountered. The feature weights are mapped with the energy of electrons. The positions of features where the transition occurs between positive and negative features are used in calculating Free-Bound transition, whereas the negative to negative feature transition is identified as Bound-Bound transition. Free-Bound transitions are considered as positive or negative depending on the type of the feature gain. Bellow is equation five for reference.

$$Free - Free\ transition = \frac{(n-1)\sum_{i=1}^{n} Free - Free\ Feature\ Gain}{Mean\ Distance}$$

Data Quality

Data quality assessment is widely mentioned in literature as one of the most important uses of provenance. Ceruti et al. even argue that a computational quality model should be an integral part of a provenance framework. Lynch advocates the integration of trust and provenance into information retrieval systems but does not discuss how provenance can be used. Li Ding et al. argue that the “where”, “who”, “why” provenance are crucial for determining the

trustworthiness of messages; however, they only develop metrics based on who-provenance (i.e., the creator of data). The research conducted by Fox and Huang introduces knowledge provenance to create an approach to determining the origin and validity of web information. In addition to using who-provenance to determine data validity, the researchers also consider information dependency when accounting for the trustworthiness of derived propositions and temporal factors when the truth value of web propositions may change over time. Prat and Madnick proposed a framework for estimating the believability of Wikipedia data based on provenance. They adopted a metric developed by Ballou, et al. for determining the temporal believability of data based on provenance information such as when the data is created. They also developed measures for deriving the believability of output data from that of its inputs based on data quality research.

Audit Trail

Various e-science applications track data provenance in the form of a workflow for scientists to verify the correctness of their own experiment, or to review the correctness of their peers' work. However, there is only one significant study that proposes a systematic approach to validating e-science experiments using provenance. Miles et al. developed a system for performing workflow validation based on provenance. According to, each workflow consists of a list of activities, and the details of these activities are recorded as provenance information in the provenance store. The system performs semantic reasoning over the properties of each activity to determine the validity of each activity. If all activities are proved to be valid, then the experiment is valid.

Replication Recipes

Zhao et al. call all the aspects of the procedure or workflow used to create a data object the “recipe” for creating that data. Obviously, it is possible to repeat the data creation or transformation if the provenance is detailed enough with precise information on each activity carried, the parameters of the activity, and datasets passed to the activity. The derivation may be repeated to maintain the currency of derived data when then source data changes or if the processing modules were modified. According to, tracking data lineage is related to the well-known view update problem. When data in one database are views derived from underlying source tables, data provenance enables the users to identify the source of the data and update it when the source data changes.

Attribution

Although there is no significant research that has been conducted specifically on this application of data provenance, it has been well recognized that “a chain of owners” is an important part of data provenance. Users can identify the creator or owner of data and verify it’s copyright. Also, data provenance acts as one form of citation when publishing scientific datasets to public databases such as GenBank and SWISS-PROT.

Analysis

We focus our analysis on data quality, since significant research has been conducted on using provenance to evaluate data. Data quality is a well established research field. Previous research on data provenance such as developed quality model framework by identifying various data quality dimensions such as data accuracy, currency, believability, etc. Many of these

dimensions such as currency and believability are related to provenance. Prat and Madnick developed a framework of data believability based on existing data quality research and define some quality metrics using bits and pieces of provenance information. Although preliminary and not comprehensive enough, this research points out a promising direction for future research. It is necessary to develop a framework for mapping various aspects of provenance (e.g. the source of data or the time of data creation) to relevant quality dimensions and design a methodology for determining data quality based on data provenance in a systematic way.

Developing software tools

In recent years, a lot of effort has been devoted to developing software tools that enable the capture and representation of data provenance. These efforts are concentrated on scientific workflows. For instance, the Karma provenance framework provides a means to collect workflow, process and the provenance of data generated from scientific workflows. Early workflow systems (e.g., Taverna and Kepler) have also been extended to capture provenance. However, the existing software systems are proprietary, relying on their own provenance models that differ in many ways. As discussed previously, provenance is becoming increasingly more important as new technologies such as Grids and web services have enabled people to engage in large-scale collaborative projects and share large amounts of data across organizational and system boundaries. Data sharing demands the sharing of its provenance. However, the existing provenance systems that rely on application-specific provenance models have made the exchange and sharing of data provenance difficult. The W7 model provides a standardization of the provenance semantics and thus helps tackle the interoperability issue among provenance models. It has been shown that it is possible to extend the W7 model to capture diverse domain-

specific provenance requirements. Study of Ram yields more than a generic ontology of provenance. A PROvenance Management System (PROMS) was developed, so that people in different domains can adopt and adapt to build domain ontologies based on the W7 model. It also provides functions for storing, browsing and querying data provenance.

Tertiary Source

As a tertiary source, encyclopedias in general could be called “irrelevant and misleading” but for the fact that their authors are trusted as having seen or studied firsthand the material about which they write. In other words, encyclopedias are accepted as legitimate sources of information largely because they have shown themselves to be useful and accurate over time and have developed a level of trust in their authors, editors, creation and publication. The example of the Piltdown man, however, should cast some doubt over the tendency toward unfailing belief in the printed word and encyclopedic knowledge in particular. Of course, such extreme examples are rare. One of the more important differences between traditional encyclopedias, such as Britannica, and a collaborative, digital encyclopedia such as Wikipedia is the issue of authorship. Modern encyclopedias exercise great control over the editorial process and use highly qualified and vetted authors that results in generally accurate and authoritative information and is largely the reason they have become well accepted and trusted sources, but this process also ensures a fairly slow development of content (Cross, 2006).

Following this tradition, Wikipedia also began using only expert authors. Originally called Nupedia, its articles were to be written by qualified and vetted authors and subjected to a high level of oversight. This ultimately proved to be a failure and Wikipedia, as it came to be called, achieved very rapid evolution and expansion by allowing anyone to generate and edit

articles – a change that opened the door to criticism over the lack of authority and quality control and contributed to the departure of cofounder Larry Sanger (Sanger, 2004) and his later develop of Citizendium, a wiki-based encyclopedia that requires contributors to use their real name and employs a high degree of oversight similar to Nupedia's original intent (Rosenzweig, 2006). A few highly publicized incidents such as the claim that former USA Today Editor John Seigenthaler Sr. was connected with the assassinations of President John F. Kennedy and Senator Robert F. Kennedy (Helm, 2005; Seigenthaler, 2005; Survey, 2006) helped fuel criticism and increase awareness of the issue among the larger public.

Despite these concerns, anecdotal evidence suggests modern users of Wikipedia generally find the content to be accurate, in-depth and usable, suggesting the model of self-governance and collaboratively constructed information is, to some extent, effective. Nevertheless, the question of authorship and article quality or overall legitimacy will undoubtedly remain as long as Wikipedia continues to operate as an open platform. These issues, coupled with Wikipedia's ease of access and frequent use by students, which could also apply to web content in general, has caused some concern among educators who feel it is not an appropriate educational tool – particularly for students who may lack sufficient background knowledge and sophistication to discern between accurate and inaccurate information. Prior to the Internet, there was less need to teach students how to determine if information was legitimate. Printed materials, which are subjected to an editorial process and peer review, were generally considered reliable sources of information. The rapid growth of the Internet, however, has created new issues. Web content does not go through the editorial process to which books, magazines and newspapers are not subjected, nor are it reviewed and filtered by librarians or teachers before being accessible to students.

European Network on Information Literacy

European Network in Information Literacy was established in 2001 with the goal of educating an information literate public. The idea of opening a European discourse on Information Literacy emerged from an initiative in USA (Presidential Committee on Information Literacy, commissioned in 1989). According to the goals of the committee, “it is important to be information literate, a person must be able to recognize when information is needed and have the ability to locate, evaluate, and use effectively the needed information” (American Library Association, 1989). It was not long before the Internet and the availability of web-based content gave new urgency to these words. The rapid growth of web accessible information and the need to be able to efficiently find it gave rise to companies such as Google and their “mission to organize a seemingly infinite amount of information on the web” (Google, n.d.). Pringle (2009) noted “the Net is an astonishing boon to humanity, gathering up and concentrating information and ideas that were once scattered so broadly around the world that hardly anyone could profit from them.” However, the process of gathering up, concentrating and organizing content simply assists in location and tells one nothing about whether or not such content is legitimate or accurate.

Jimmy Wales, founder of Wikipedia, had a different goal – “a world in which every single person is given free access to the sum of all human knowledge” (as quoted in Lih, 2009). Although not specifically addressed in Wales’ comment, the “sum of human knowledge” would necessarily, one would assume, need to be legitimate and reliable information. Early efforts to use qualified and vetted authors were unsuccessful (Lih, 2009; Rosenzweig, 2006) and in order to accomplish their goal, Wikipedia adopted an open editing process that allowed anyone to

participate. While this decision proved to be highly successful with Wikipedia growing from just a few hundred articles in 2001 to over 3.7 million by 2011 and 50 times the size of the next largest English language encyclopedia ("Wikipedia: Size comparisons," n.d.) it also gave rise to concerns over the accuracy, authority, and overall legitimacy of the content. I initially had my own concerns over the use of Wikipedia in academic circles, but, as I watched it grow and found myself using it more and more, I realized that students needed to learn to determine the legitimacy of Wikipedia content, and web content in general, for themselves – particularly because it was clear they were using it more and more.

I have observed that students' approach to web-based content, including Wikipedia, often paralleled Freire's (2000) model in that they saw information as external and disconnected from themselves, the words of apparent experts that could not, should not, be questioned. This is undoubtedly due, in part, to the banking model (Fieire, 2000) of education that has as its focus the filling of students' heads with facts of the world for later withdrawal – often in the form of a test of their memory and retrieval skills. The analytical and critical approach to learning has often been overlooked. However, as Temple (2005) points out, "only those whose critical faculties have been nurtured, through dialogue about the issues that matter in their lives, develop critical consciousness" (p. 16). Wikipedia actually offers a unique opportunity to teach students to doubt, question, analyze and explore the legitimacy of apparent factual claims and encourage their development of critical literacy and critical consciousness. Drawing on an idea presented by Harouni (2009), I created an online poll, asking respondents to select an article in Wikipedia about which they felt they already knew something or considered themselves an expert and then read the article taking note of anything they found that they did not agree with or trust. They then had to verify whether or not this suspect information in Wikipedia was correct.

There was one respondent who was reading an article on the Denver Broncos football team and felt the information regarding the Broncos only having two NFL Hall of Fame members was surely wrong. In order to verify his suspicion he went to the source of the information – the National Football Hall of Fame. He discovered much to his dismay, that at that time (early 2011) the Denver Broncos did in fact have only two Hall of Fame members. Others discovered that the origin of the Australian Shepherd is convoluted and may have little to do with Australia or that the manner of Hitler's death is in dispute and relies somewhat on whose testimony you chose to believe. This type of research was played out over and over as respondents identified suspicious information, at least to them, and then went through the process of verifying it. The results were illuminating. Most respondents commented that they were surprised to find that "Wikipedia is usually right" and wondered why they had been repeatedly told by educators that it was not reliable.

Others noted that while the information was not wrong, it was often incomplete or had simplified a more complex issue, such as the origin of the Australian Shepherd, into a sentence or two that obscured a deeper issue. Perhaps due to years of persuation by former instructors on the evils of Wikipedia, a few respondets continued to maintain that Wikipedia was often wrong and full of errors. Further questioning, however, showed that these respondents tended to hold on to misconceptions or were unsuccessful in finding alternative sources of information and chose to simply believe themselves correct. While such sketches are interesting, they do not provide any assurances needed regarding the overall legitimacy of Wikipedia and other web based content nor do they fully develop the skills necessary for an information literate society. It is also important to remember that Wikipedia is only one example, even though large and popular, of collaboratively constructed knowledge. Wikis exist all over the web for a variety of purposes and

educators are finding the collaborative nature of the wiki a powerful educational tool that supports the development of 21st Century Skills including communication, collaboration, problem solving, critical thinking, knowledge construction, and participation in a global community (International Society for Technology in Education, 2007). In my own experience, wikis have proven to be a unique educational tool.

Collaborative nature

The collaborative nature of the wiki allows anyone to contribute to a single shared database. Furthermore, the wiki is not a static, single use product but a living document that can be added to each year while preserving data from prior years. As this collection of data grows, users can perform different types of analyses depending on their information needs. Wikis are also used to share information on any number of individual topics or projects. Software projects often offer some sort of online documentation for users and the wiki is a perfect tool for both developing the documentation and providing access.

History of the Encyclopedia

According to the Encyclopedia Britannica (2002), the term “encyclopedia” comes from the Greek words enkyklios paideia meaning well-rounded or general education, or the circle of learning (Kister, 1994; Kogan, 1958), and the modern encyclopedia is a realization of this implied intent (Collison, 1966) – a book or collection of volumes that “contains information on all branches of knowledge” (“Encyclopedia,” 2002), or, as Thoreau (1910) put it, “an abstract of human knowledge” (p. 195). In his Naturalis Historia (79 CE), Pliny the Elder used these words to describe the content of his work as containing the circle of Greek learning (Kogan, 1958;

Stockwell, 2000). Stockwell contends that it was not until 1531 when these two words were combined in the term “encyclopedia” by Sir Thomas Elyot in his Bok of the Governor, or, according to Kister (1994) in the title of the Latin work Encyclopedia: seu, Orbis Disciplinarium, tam Sacrarum quam Prophanum Epistemon published in 1559 by Paul Scalich. Despite their long history, dating back at least to the fourth century B.C. (see Collison, 1966 for an extensive chronology), and importance, Thorndike suggested they are “the most important monuments of the history of science and civilization” (1924, as cited in Kafker, 1981), the encyclopedia has not been greatly studied (Kafker, 1981).

Nevertheless, the encyclopedia has a rich history dating back to the ancient Greeks. Collison (1966) considered Plato to be the father of the encyclopedia. Although Plato never wrote an encyclopedia himself, he was the founder of the Academy of Athens and was also uncle and mentor to Speusippos who did compile an encyclopedia based on the teaching of Plato to use in his own teaching. One of the earliest known attempts at creating a vast compendium of knowledge is the Naturalis Historia of Pliny the Elder (77 C.E.). His thirty-seven books attempted to cover the known natural world and included over 2,500 chapters on topics such as “geography, physiology, zoology, botany, and medicine” (Kister, 1994, p. 5), and, similar to the modern encyclopedia, compiled information from two thousand works and over four hundred authors (Kogan, 1958; Lih, 2009). The Chinese T’ai P’ing Yu Tan, published in the tenth century, is generally considered the first modern encyclopedia (Kogan, 1958). The first work to be titled “Cyclopaedia” was compiled in 1541 by Ringelberg (Kogan, 1958). The father of the modern encyclopedia, however, is probably Ephraim Chambers who published the two volumes Cyclopaedia: or, An Universal Dictionary of Arts and Sciences in London in 1728 which

introduced now common elements such as alphabetical arrangement and included a system of cross-references (Kogan, 1958; Lih, 2009).

The most comprehensive early encyclopedia was undoubtedly Diderot's much larger, eventually comprising 28 volumes, French Encyclopedia published between 1751 and 1772. Originally intended as a translation of Chamber's Cyclopaedia, it abandoned the impartial and objective (Kister, 1994) point of view and focus on sharing general knowledge of earlier (and later) encyclopedic efforts, and instead presented its own point of view and even commentary on the state of France and Europe which resulted in attempts at censorship, confiscation by police, orders to have copies burned, and Diderot eventually having to work in secret in order to finish (Kogan, 1958). The first truly comprehensive English language work is generally considered to be The Encyclopedia Britannica originally published in weekly installments beginning in 1768 ("Encyclopedia," 2002; Kister, 1994; Kogan, 1958; Lih, 2009) and repeatedly in fourteen subsequent editions – the most recent of which was published in 2002.

Encyclopedia Britannica

Encyclopedia Britannica claims that it has "evolved into the largest and most comprehensive general encyclopedia in the English language ("Encyclopedia," 2002). Despite their attempt at being a general work of knowledge for common people ("Encyclopedia," 2002) and "accessible, both physically and intellectually, to students and other users in as fair, accurate, and precise a manner as possible" (Kister, 1994, p. 3), the encyclopedia has not been readily accessible to average users due to its rather large size and expense (Kogan, 1958). In 1938, H. G. Wells, in arguing for a world encyclopedia pointed out that encyclopedias had largely been reserved for only an elite minority. Even today, users generally have to visit a local public or

school library to use an up-to-date encyclopedia. While newer encyclopedias, such as The World Book Encyclopedia first published in 1917, attempted to be more family oriented, using glossy pages and color illustrations, the encyclopedia has never become a common addition to home libraries (Lih, 2009). Furthermore, due to continually evolving content, anyone who manages to purchase an encyclopedia will also find their expensive investment increasingly out of date; a problem which likely limits the number of non-institutional owners.

Although Wells did not specifically mention an electronic encyclopedia, shortly thereafter, Vannevar Bush (1945) proposed what may well have been the precursor to hypertext and digital content. In laying out the foundation of his Memex, Bush focused on the power of "associative indexing... whereby any item may be caused at will to select immediately and automatically another." Ultimately, he envisioned that the Memex would give rise to "wholly new forms of encyclopedias." While the Memex never saw the light of day, the advent of the personal computer did give rise to new forms of encyclopedias stored on optical media. In 1993 Microsoft Corporation released Encarta on CD-ROM. While not overly impressive, copies were often included for free in the purchase of new computer, it was often sufficient for home users (Lih, 2009). For the first time, average home users had ready access to encyclopedic content. Microsoft continued to improve its product and Britannica released their own electronic version in 1994 – for $995 (Lih, 2009). The rapid growth of the Internet, however, began to undermine the usefulness of CD-ROM-based encyclopedias – especially because all major players were moving toward online, subscription-based content. Seekers of information, however, found that a quick search of the Internet was becoming an effective tool for finding information and was cheaper and even faster than loading a CD-ROM or setting up a subscription.

Unfortunately, such ease of access was putting users at odds with credible and legitimate information. The Internet may have become the ultimate realization of Bush's Memex, but instead of being deliberately filled with the collected works of humanity it was largely a playground in which anyone could post anything at any time without any sort of editorial or peer oversight. By 2000, the Internet was a wellspring of information but with increasingly divergent and competing purposes. However, in 2001, the advent of Wikipedia began to change the landscape of information seeking on the Internet. Wikipedia might be considered a necessary outcome of technological progression. Individuals such as Bush, McLuhan, and Wells all hinted at various capabilities that have combined in the form of a large, collaborative collection of human understanding. One wonders if Wales had not begun Wikipedia if someone else eventually would have begun something similar. Most people know Wikipedia by what it is today – a vast, free, online encyclopedia freely accessible and editable by anyone (see figure 2). However, that is not how it started. In his book The Wikipedia Revolution: How a Bunch of Nobodies Created the World's Greatest Encyclopedia, Lih (2009) details how this came to be. According to Lih, Wikipedia began as a very tightly controlled project called Nupedia. Unlike its successor, Nupedia had a very convoluted process of article development.

While the initial project did rely on volunteers from the start, in order to maintain integrity, authors and editors had to be carefully vetted and either hold a doctorate or otherwise be a recognized expert in their field, and each article would go through a lengthy seven-step process to ensure integrity. The process, however, proved to be too time consuming with only tens of articles produced in the first year (Rosenzweig, 2006; Lih, 2009). Wikipedia was made possible largely due the work of Cunningham (Leuf & Cunningham, 2001) who developed the idea and implementation of wiki software which he called the wikiwikiweb from the Hawaiian

word wiki meaning fast (Kane & Fichman, 2009; Lih, 2009). Simply put, a wiki is a website that can be edited by anyone (Kane & Fichman, 2009), in the case of one that does not require registration, or only by members of a particular wiki or community. The initial iteration of wikiwikiweb was released in March of 1995.3 Wales and co-developer Sanger eventually became aware of the wiki software and in an attempt to accelerate the slow pace of article development on Nupedia set up a variation of the original wiki software called UseModWiki which ran on a web server in January 2001.

Editing process

Although it generated interest, it also was criticized for its open editing process that was counter to the initial intent of Nupedia and a week later it was moved to wikipedia.com to continue the experiment. At that time it was still seen as part of the Nupedia project and articles developed there were to eventually be moved to Nupedia (Lih, 2009). While ultimately a failure, the founding principles of Nupedia survived and ultimately gave rise to what is easily the world's largest encyclopedia (Rosenzweig, 2006). Nupedia took its name from the GNU Manifesto written by Richard Stallman in 1985. The manifesto laid out the ground work for the free software movement which had at its core the idea of freedom, that software users had the freedom to examine, modify and redistribute software to suit their needs. An important element of the GNU manifesto was that users not only had the right to redistribute software, they had the obligation to share back their changes and could not restrict the rights of future users to also examine, modify and redistribute (Stallman, 1985). These principals are at the core of Wikipedia which encourages users to modify and redistribute content.

Wikipedia has grown to be one of the most popular sites on the web. Worldwide, according to Alexa statistics October, 2011) the fifth most Wikipedia has ranked as high as media tools continue to grow popular sites as the interests of Internet users are continually in flux. It is likely that Wikipedia will continue to be a highly trafficked site. As evidenced by its high ranking, top search positions as well as top positions in global traffic, it is not surprising that current growth is relatively low as a large percentage of Internet users are already visiting Wikipedia. Additionally, Wikipedia's article count continues to grow as well and currently contains over 3.7 million articles in English alone and 20 million, as of November 2011, in all languages combined ("Wikipedia:Size comparisons," n.d.). Similar to traffic patterns, article growth rates have fallen off in the past couple of years after exponential growth between 2005 and 2010 when it grew from approximately 500,000 to over 3 million ("History of Wikipedia," n.d.). This is likely due to the decreasing number of potential topics yet to be included.

Research on Wikipedia

Despite its popularity, Wikipedia receives a steady stream of criticism regarding its overall reliability and credibility (Emigh and Herring, 2005; Giles, 2005; Rector, 2008; Rosenzweig, 2006). Not surprisingly, former Britannica editor-in-chief Robert McHenry has been a vocal critic focusing on the open editing process that ensures constant change but no guarantee of improvement and places more importance on being free than it does on being reliable. He states, somewhat humorously, the user who visits Wikipedia to learn about some subject, to confirm some matter of fact, is rather in the position of a visitor to a public restroom. It may be obviously dirty, so that he knows to exercise great care, or it may seem fairly clean, so that he may be mitigated into a false sense of security. What he certainly does not know is who

has used the facilities before him. (McHenry, 2004) One of the most widely reported events that called Wikipedia into question was the creation of a biography linking former USA Today Editor John Seigenthaler Sr. with the assassinations of President John F. Kennedy and Senator Robert F. Kennedy (Helm, 2005; Survey, 2006).

Seigenthaler (2005) himself denounced the entry stating “I have no idea whose sick mind conceived the false, malicious ‘biography’ that appeared under my name for 132 days on Wikipedia, the popular, online, free encyclopedia whose authors are unknown and virtually untraceable.” Wikipedia does not ignore such concerns and criticisms and even maintains an article on its own reliability (“Reliability of Wikipedia,” n.d.). However, Wikipedia has achieved its phenomenal growth primarily because it opened up its editorial process to anyone and it now has approximately 3.7 million articles in English written by anonymous authors compared to Encyclopedia Britannica's 65,000 articles in print or 120,000 articles online (Berinstein, 2006) written by their 4,800 worldwide, paid contributors (according to Tom Panelas, director of corporate communications at Britannica as quoted in Berinstein, 2006). Despite criticisms, there have been a number of studies suggesting that Wikipedia is fairly reliable. The often cited study in Nature (Giles, 2005), found errors in Britannica and Wikipedia. Their review of 42 science articles by content experts found only eight serious errors, defined as misrepresentations of important concepts, which were evenly split among both Wikipedia and Britannica. The study also found 162 factual errors or misleading statements in the Wikipedia articles and 123 in Britannica or an average of four in each Wikipedia article and three for Britannica - a difference they described as “not particularly great” (Giles, 2005).

However, Internet skeptic and author of The Shallows: What the Internet is Doing to Our Brains, Carr (2006) noted that a more in-depth review of the study showed that it “probably

exaggerated Wikipedia's overall quality considerably." Furthermore, after conducting his own review of the study, Carr summed it up consequently: If you were to state the conclusion of the Nature survey accurately, then, the most you could say is something like this: "If you only look at scientific topics, if you ignore the structure and clarity of the writing, and if you treat all inaccuracies as equivalent, then you would still find that Wikipedia has about 32% more errors and omissions than Encyclopedia Britannica." That's hardly a ringing endorsement. Fortunately, other studies of Wikipedia have been conducted. With respect to perceived credibility, Chesney (2006) studied the perceptions of subject experts and nonexperts on a variety of Wikipedia articles. A total of 258 academics (defined as research fellows, research assistants and doctoral students) were surveyed (with a 21 percent completion rate) and randomly given either an article in their own area of expertise or a random article and asked to review and assess the credibility of the article, the authors and Wikipedia in general.

While both groups did not differ in their assessments of author and site level credibility, there was a significant difference in perceived credibility of articles with the subject experts rating articles more credible than the non-expert, random assignment group – suggesting a high level of accuracy in Wikipedia (Chesney, 2006). It was noted, however, that experts found errors in 13 percent of the articles which is consistent with the findings of others (Giles, 2005; Rector, 2008). Rosenzweig (2006) also found slightly more errors in Wikipedia than comparable reference works but also pointed out they were minor. Rector (2008) found that Wikipedia was less accurate than other sources (80% accuracy compared to 96% in Britannica). In other words, while errors persist in Wikipedia and in more traditional encyclopedias, such as Britannica, there is still a fairly high degree of accuracy and perceived credibility in Wikipedia. Precisely why

non-experts felt articles were less credible (Chesney, 2006) was not directly addressed; although it is possible that non-experts lack sufficient background to accurately judge an article.

Legitimacy

However, because it is reasonable to expect that many users of Wikipedia would be non-experts, providing a means by which such users can judge the legitimacy of content would be beneficial. Magnus (2006) conducted a similar study in which copies of articles of similar depth in both Britannica and Wikipedia were given to experts for a blind review. The study used a small sample of three articles on somewhat obscure topics: Rawls' theory of justice, Husserl and phenomenology, and bioethics. Experts differed in their evaluations of the articles. The Wikipedia article on bioethics was called bizarre and not written by someone in the field. However, a reader of Husserl called the Wikipedia entry his favorite adding that it was how an encyclopedia article should be written. Magnus (2006) noted that variability in the quality of Wikipedia articles "should come as no surprise, since Wikipedia entries rely on contributors. Different entries will attract contributors" (p. 4). Others (Halavais, 2004 as cited in Read, 2006; Magnus, 2008) have attempted to track the longevity of errors they inserted themselves with varying results. It should be noted that intentionally inserting errors in Wikipedia is considered vandalism and discouraged (Kane & Fichman, 2009). Magnus (2006) pointed out that Wikipedia articles change over time and evaluations of old articles do not inform us about the content of newer versions. He suggested we need ways of evaluating changes in Wikipedia over time.

A time-based approach to evaluating the accuracy of Wikipedia was conducted by Luyt, Aaron, Thian & Hong (2008) who focused on the age of edits. For their study, the authors selected the same 42 articles used in Giles (2005). The earlier study included information on the

exact errors that reviewers found which allowed Luyt et al. (2008) to pinpoint the versions of the Wikipedia articles where the errors were introduced. This was accomplished using the history feature of Wikipedia that preserves every edit with a time and date stamp as well as the name of the user or IP address responsible for the edit. They referred to this process as assigning blame, and tracked the existence of each error in terms of total number of edits between the introduction of the error and its removal, and the overall amount of time in days between the introduction of the error and the time of the review in Giles (2005). The purpose of the study was to test Cross' (2006) theory that older information that has withstood the test of time would be more accurate and that errors would be attributable to more recent edits that have not had the opportunity to be fully examined. Luyt et al. (2008) found no support for this theory instead finding that at least 20 percent of errors could be attributed to the initial edit that began the article which they called a "first-mover" effect. They concluded that attempts to validate Wikipedia content based on the age of the surviving edits would be unable to accurately account for this first-mover effect.

The implication for Wikipedia and its users is that metrics such as edit age and article maturity are not going to be usable as a tool to measure accuracy or legitimize Wikipedia content. Researchers have also attempted to evaluate the verifiability of Wikipedia articles by looking at citations. Luyt and Tan (2010) randomly sampled 50 history articles from Wikipedia and compared the citations in those articles with citations from articles in the Journal of World History (JWH). In the 50 Wikipedia articles they found a total of 508 citations of 480 distinct references. The 18 articles from JWH, by comparison, contained 1,877 citations of 1,351 distinct references. When comparing the types of references cited, they found 62 percent of Wikipedia citations were of Internet sources compared to 1.2 percent for JWH. Such results, they suggest, indicate that Wikipedia is reliant on low level, non-academic sources of information.

Supporting Evidence

Whether or not such comparisons are fair is another issue. Scholarly journals exist for an entirely different purpose than encyclopedias, and attempts by Wikipedia to add supporting evidence should be encouraged. Furthermore, scholarly journals tend to focus on original research, which is held to a high standard and expected citation practices. Reporting of original research is specifically prohibited in Wikipedia ("Wikipedia: No original research," n.d.) as it is primarily focused on providing information on general knowledge for common people similar to printed encyclopedias. Other approaches to evaluating Wikipedia and wikis in general (such as those used in business or the classroom), focus on measuring and evaluating editor contributions. Arazy et al. (2010) proposed a new set of algorithms to calculate authorship in wikis.

They pointed out that previous methods to calculate author contributions tended to be flawed due to their focus on basic metrics automatically tracked by wikis such as the number of page edits for each unique contributor – WikiDashboard4 being one such tool. Other attempts focused on evaluating a user's contribution by comparing a current version to a previous one for a particular user's contributions with the sum of all contributions providing a measure of a user's overall effort (Hess, Kerr and Rickards, 2006 as cited in Arazy et al.). Still other approaches mirror efforts currently under investigation by Wikipedia such as measuring the longevity of edits (Adler, de Alfaro, Pye & Raman, 2008; Cross, 2006, Luyt et al., 2008) which is similar to a color-coding scheme currently being explored (Claburn, 2009; Cross, 2006; Leggett, 2009), and the use of a rating system to calculate a user's reputation and, by http://wikidashboard.appspot.com/ extension, their overall level of contribution (Sabel, 2007). Key differences exist between Sabel's approach and the one currently being explored by

Wikipedia (“Wikipedia: Article feedback tool,” n.d.). Sabel’s (2007) approach proposes weighting the similarity of page versions and assigning an adoption coefficient which can then be used as part of a reputation system which could function as a measure of overall contributions and reliability.

Wikipedia’s implementation, part of an overall strategic plan (“Strategic Plan/Movement Priorities,” n.d.), have readers rate articles on four criteria: trustworthy, objective, complete, and well-written. There is also a box for readers to check if they are “highly knowledgeable about this topic.” It is interesting; however, that Wikipedia does not view this feedback tool as a measure of quality or accuracy. Of the tool, Wikipedia states, the current version of the tool represents a starting point. The Wikimedia Foundation wants to encourage direct reader engagement as a good way to quickly acquire qualitative feedback and to make more readers aware that they can directly improve Wikipedia. We hope that this tool will help the readers in the Wikipedia community become active editors. Less knowledgeable users, however, are likely to view an article with a high rating as a more trustworthy or objective article than one with a lower rating regardless of the overall intent. Furthermore, it is unclear how the Wikipedia article feedback tool would account for vandalism or the inevitable changes in articles over time. Contrary to these approaches, Arazy et al. (2010) propose a new approach for calculating editor contributions to wikis by first breaking edits types into five categories (add, improve navigation, delete, proofread, and adding links) and measuring contributions in each category.

Quantity of contributions

They focused on the quantity of contributions and not the quality which they considered quite difficult to measure. They also suggested longevity could be used as a quality measure

because the evolution of wiki pages should involve the removal of low quality content while allowing high quality content to remain. Similar to Luyt et al. (2008), Arazy et al. (2010) failed to find support for this premise. Precisely why errors tend to linger has not been addressed. However, it is possible that errors not addressed within a certain amount of time tend to gain a certain level of legitimacy and may be overlooked by all but the most diligent and knowledgeable editors. To test their approach, Arazy et al. (2010) compared their algorithms against nine randomly selected and human scored articles in Wikipedia. They found a high level of correspondence between their algorithms and human scores. The results were then used to create visualizations of editor contributions across the five categories.

This resulted in several different glyphs showing relative percentage of contributions for editors and is intended to be included on the corresponding article page. These were then user tested to determine their effectiveness. They note, however, that this is contrary to the collaborative and unattributed nature of wikis, but see potential application in classroom or research settings as a way to increase motivation and participation. Teachers using wikis as class projects could also benefit from having a way to evaluate the work of individual members of a group. The value of such visualizations in Wikipedia itself are uncertain because knowing which users contributed in which way does not help us to know if those users are knowledgeable or credible. Glyphs or similar visualizations, however, could potentially be used to provide a form of feedback on articles and how they are related to other articles via the patterns of the contributors. Algorithms such as those developed by Arazy et al. (2010) could prove useful in calculating and visualizing such relationships. Knowing who the major contributors are to individual articles may also be useful in evaluating content if one could track and measure their contributions across Wikipedia.

Similar to Raymond's (1998) comment regarding Open Source software development, that "given enough eyeballs, all bugs are shallow," with enough editors, Wikipedia articles are potentially more credible and accurate. A glyph similar to the one suggested by Arazy et al. (2010) could be used by visitors to Wikipedia to easily visualize if an article was mostly written by many editors or just a few and if the edit history of those editors supports an authoritative background or not. Other studies have focused on comparisons between Wikipedia articles and professionally maintained information stores. In their study of the accuracy of cancer information on Wikipedia, Rajagopalan et al. (2010) chose 10 articles on types of cancer to compare with the information on a professionally maintained database, the National Cancer Institute's Physician Data Query (PDQ) cancer database. With respect to Wikipedia they found that errors were rare (less than 2%). The Wikipedia articles were also found to be less readable than those on the PDQ database. Interestingly, this readability was measured using the Flesch-Kincaid grade-level scale which found a grade level score of 9.6 for the PDQ database and 14.1 for Wikipedia (higher numbers are considered less readable). This could also be interpreted as meaning that the Wikipedia articles were written at a higher level, as would be assumed from more knowledgeable authors.

They also found no significant difference between the depths of coverage of Wikipedia articles compared to the PDQ database. More recently, Arazy et al. (2011) attempted to measure how several factors, cognitive diversity, group member orientation (administrative or content), and task conflict, interact and what effect they have on the quality of information in Wikipedia. The study used a stratified sampling approach that randomly selected 15-17 articles from six of Wikipedia's top-level categories: culture, art and religion; math, science, and technology; geography and places; people and self; society; and history and events. They sampled a total of

96 articles using Wikipedia's random article feature. A unique aspect of the study was the focus on cognitive diversity. They argued that deep-level diversity, which relates to education, expertise and knowledge, can enhance groups' performance, especially when the task is cognitively complex and requires multiple perspectives or entails creativity, since cognitive diversity increases the variety of perspectives brought to a problem, creates opportunities for knowledge sharing and leads to greater creativity.

When looking at diversity on a per article level, they found a very high level which suggests very little overlap in the activity of the contributors outside the current article. Article quality was measured using independent ratings by senior librarians at a large North American university followed by a negotiated consensus to arrive at a rating. Although article quality was not the primary focus of the study, rather the extent to which group characteristics influenced quality, they nevertheless found that article quality was moderately high scoring 4.4 on a 7 point scale. Despite indications that Wikipedia is an often accurate and credible resource, concern over who writes the articles continues. The notion of authorship is deeply ingrained in the process of writing, citation and our overall judgement of authority and credibility. In major publication style guidelines, such as the American Psychological Association (APA) style, the Modern Language Association of America (MLA) style, The Chicago Manual of Style (CMOS) and others, prominence is placed on the author of a work. Such citations follow an author, year (APA, 2001), or author, page numbers (MLA, 2008) format, but what is consistent is the focus on the author. Early encyclopedias, such as the Naturalis Historia of Pliny the Elder (77 C.E.) also considered the author as primary. Pliny referenced 473 mostly Greek authors in his 2,493 articles (Stockwell, 2000).

Authorship

The role of authorship, however, has historically not been a constant. As Foucault (1984) points out in his essay "What is an Author," the importance of knowing the author of a text has changed over time. Text that we would now tend to classify as literary were at one time accepted and passed along without concern over knowing the author, while in the middle ages, scientific texts were generally only accepted as true when attributed to their author. The modern approach has more or less reversed the importance of Foucault's author function. Modern scientific discourse places little emphasis on the author while we place great importance on the author of literary texts. There is, for example, some debate over the true author or co-authorship of Shakespeare's works (Foster, 1999; Vickers, 2004) even though knowing the name of the author will not change the nature of those texts but could, if we can prove that it was not Shakespeare, change how they are received. Conversely, finding out that Einstein did not develop the Theory of Relativity would likely have little impact on the nature of that discovery and its use and importance in various scientific fields though it might change our perceptions of Einstein. Interestingly, Foucault does make exception for the few individuals who have essentially made certain discourses possible – what Foucault called "founders of discursivity" (p. 114).

Foucault identifies Freud and Marx as examples of individuals who not only wrote their own works but also opened the door to endless possible discourse such as Freudian psychology or Marxism. That, too, may be changing as Einstein's Theory of Relativity is now so widely accepted and intertwined in various scientific fields that it is often referred to as simply relativity, without reference to Einstein. For example, "another prediction of general relativity is that time should appear slower near a massive body like earth" (Hawking, 1988, p. 32). More recent conversations on Communism and Socialism rarely reference Marx unless it is to point out

discrepancies between modern implementations and Marx' original intents. Modern encyclopedias, however, continue to place traditional importance on the author. Both the Encyclopedia Britannica and the World Book Encyclopedia put authors of articles in the headlines. Of its contributors, the Encyclopedia Britannica states, to meet these challenges and opportunities, Britannica has done what we have always done throughout our 240-year history: sought the very best minds in the world to help us.

In the past, they had names like Albert Einstein, Sigmund Freud, Marie Curie, Bertrand Russell, T.H. Huxley, and George Bernard Shaw, all of whom were Britannica contributors in their day. (Encyclopædia Britannica Board of Editors, 2010) Wikipedia, conversely, takes the opposite approach and relies not on the credibility and recognition of its authors but on citation and the verifiability of its content ("Wikipedia: Verifiability," n.d.) as well as an informal form of peer review inherent in socially constructed knowledge or the wisdom of the crowds (Arazy, Morgan, & Patterson, 2006; Surowiecki, 2005). The extent to which it is achieving that goal is debatable, but the shift in focus is not without merit. Foucault (1984) argued that while authorship was regarded as essential to "truth" in the middle ages, in the seventeenth and eighteenth centuries "scientific discourses began to be received for themselves, in the anonymity of an established or always redemonstrable truth... and not the reference to the individual who produced them" (p. 109).

In other words, scientific discussions generally exist separate from the author. Whether or not various areas of Wikipedia should be treated differently based their author function is another discussion. The current state of Wikipedia ensures we may never know the name, background, credentials, etc. of the true authors of each and every article. However, it may be possible to develop profiles of authors and articles through a process known as social network analysis.

Social Network Analysis Social network analysis (SNA) is a research methodology with the primary goal of identifying patterns of social relationships based on the connections of actors to each other (Scott, 1991; Wasserman & Faust, 1997). Haythornthwaite (1996) described SNA as "an approach and set of techniques for the study of information exchange" (p. 323).

The focus is on the "patterns of relationships between actors" and resources that can include actual goods and services as well as less tangible items such as information. Furthermore, according to Haythornthwaite (1996), the process is empirical and focuses on observable relationships, the networks, between the actors. Additionally, de Laat, Lally, Lipponen, & Simons (2007) suggested that SNA can help in "identifying patterns of relationship between people who are part of a social network" and "assist us in the analysis of these patterns by illuminating the 'flow' of information and/or other resources that are exchanged among participants" (p. 89). Only after an examination of these relationships are they grouped according to the strength of their connections to other regions of the network (Monge, 1987). Actors can also be members of more than one network based on their relationships. The patterns that develop help us understand with whom individuals interact and how they exchange information. Although developed well before the advent of computers and computer networks, SNA researchers are increasingly looking at ways to understand online networks.

Document Categorization and Wikipedia

The traditional model for document representation is a word-based vector (Bag of Words, BOW), where each dimension is associated with a term of a dictionary and represents number of occurences of the term in a document. The majority of existing text classifiers represent documents as an unordered collection of words - a bag of words (BOW). Although is the Bag of Words model powerful, simple and commonly used, it has several drawbacks. As Gabrilovich

and Markovich (2006) state, this method is very effective in up to medium difficulty categorization problems, where the category of a document can be determined by various easily recognizable keywords. Limitations of this method become more significant when applying to more challenging tasks, such as categorization of short documents or when dealing with small categories. Several studies have proposed mapping terms and phrases within documents to their corresponding articles in Wikipedia. These articles represent aggregations of common characteristics of a certain concept. Gabrilovich and Markovich (2006) as well as others (Wang & Domeniconi; 2008; Huang, Milne, Frank & Witten, 2008) proposed a way to enrich the BOW representation of processed documents using the semantic knowledge from Wikipedia, via linking terms and phrases to concepts from Wikipedia.

Gabrilovich and Markovich (2006) applied standard text classification techniques to link document texts to Wikipedia concepts. They proposed building a feature generator, which determines the most relevant concepts complementing the document and creates a set of features that augment the bag of words. Feature generation yielded considerable improvement in classification of short documents.

Wang and Domeniconi (2008) also proposed a method to overcome the shortages of BOW approach. Their approach attempts to highlight the semantic content of documents by embedding background knowledge constructed from Wikipedia into a semantic kernel, which is then used to augment the BOW representation of documents. By embedding Wikipedia-based kernels into document representation, Wang and Domeniconi were are able to keep multi-word concepts unbroken, capture the semantic closeness of synonyms and perform word sense disambiguation hence to show benefins and potential of embedding semantic knowledge and surpasses limitations of BOW model.

Janik and Kochut (2008) devised an ontology-based text categorization method using an RDF ontology derived from Wikipedia. A valuable aspect of this approach is that their method does not require a training set, being based on ontological knowledge. Moreover, this method performs almost as accurately as statistical methods trained on the documents from the categorization ontology. The efficiency of this model relies heavily on a rich and comprehensive ontology acquired from Wikipedia, an ontology that can be very well used as a text classifier.

CHAPTER 03: METHODOLOGY

Overview

The purpose of this research is to develop the understanding how document classification can be improved using Wikipedia. In this section of the report, the research methodology is outlined specifying the research approach and method that was used. The methodology for the research takes into account the knowledge obtained from the review of related academic literature, the nature of the research subject and the aim/objectives which the research wishes to achieve. The research will be carried out using an empirical research using both qualitative and quantitative research. I believe using a mixed method approach will enhance my research (Kroll, B., & Taylor, A, 2003, pp. 54).

Research Design

The Internet is considered as an important tool in obtaining relevant information needed to find a series of articles in magazines and newspapers in the database. To test the research hypothesis, the research conducted by a three step process: the construction of a research pool issue, the validation of products and piloting of articles. Moreover, data from various sources, some of which will be collected online, while some are on paper. Hard data is primarily the result of a thorough analysis of the materials you find online. The research has involved analysis of news messages over the Internet through a period of years. The method is to read the summary or body of each publication.

Descriptive Research

Sekaran (2000) defines descriptive research as a method used to make an understanding of the attributes of different individuals/corporations by allowing them to think in a systematic manner about different elements. Furthermore, this type of method is available when the information on which the research is to be carried out is easily obtainable and the researcher is much more aware about the situational factors of the study. Descriptive research is a reflection of the correct outline of an individual, his actions, procedures and situations.

Exploratory Research

According to Agrawal, A. and Mandelker, G.N. (1987 pp. 823-37) an exploratory research is conducted when the overall objective of the study is to clarify and explore the research issues. In this type of situation limited information is available to the researcher, (Sekran, 2000).

Data Collection Methods

A data collection procedure was developed which checked the value of the ideas behind the research by reviewing extensive literature. Triangular approach using different data sources and collection methods are particularly useful in research, theory generation (Orlikowski, 1993). Special attention should be paid at the time of collection of data to avoid deviations of the things learned from literature (Jeong, 2009).

Quality research

Among qualitative social research in the social sciences, the collection of non-standard data understood and their evaluation. Most frequently it is interpretive and demonstrative methods used as an analytical means. Theoretical foundations of qualitative methodologies in the social sciences provide among other theoretical traditions such as the phenomenological or symbolic interactions , often under the name of the interpretive paradigm or interpretive sociology are combined. In everyday life, and shared by scientists and non-scientists living environment and the construction of meaning are reasonable character of social action in specific cultural contexts always existed already before the sociological analysis generally turns its object. In contrast to scientific facts of social science subjects are always already pre-structured so in some way by the person examined and questioned, and thus reflexive. The traditional methods of qualitative research try to address this particular character of social science subject areas through the open character of the data collection and interpretive nature of the data analysis into account. This qualitative research than anything, if they are committed to the Interpretative Sociology, usually a high value on the acquisition of the actor's perspective, and action orientation and the interpretation of patterns of respondents.

Threats to Validity

Yin identifies threats to construct validity, internal validity, external validity and reliability as being applicable to case studies. Construct validity implies that the domain ontology of provenance may be either incorrect or incomplete. Benbasat et al. and Yin suggest that using multiple sources of evidence controls threats to construct validity. Yin also suggests that key informants review the draft case study report. In our case study, data collection and validation

interviews helped ensure construct validity. The data collection interviews helped capture the requirements. The validation interviews ensured that data collected was correct and complete. All the information collected from interviews and system review was cross-validated. Trochim refers to internal validity as credibility that better reflects the underlying assumptions for case studies. He further argues that study participants should be the ones who can legitimately judge the credibility of the results. Multiple iterations of validated interviews helped ensure internal validity of our findings. External validity deals with the problem of knowing whether a study's findings can be generalized beyond the immediate case study.

Yin suggests using an analytical generalization can help an investigator link a particular set of results to some broader theory. Trochim argues that transferability (i.e., the degree to which the results can be generalized) can be enhanced by describing the research context and the assumptions that are central to the research. We describe in detail the context and assumptions of the case study. A threat to reliability arises if a methodology will not be repeatable by other researchers. Yin suggested use of a case study protocol, i.e., a document that lists all the activities undertaken by the researcher conducting the case study. This allows other researchers to conduct similar case studies in other settings and to compare results across case studies. We described details related to the methodology earlier in this section.

CHAPTER 04: EXPERIMENTAL DESIGN

Proposed Methods

In this chapter, we will propose how can Wikipedia and the semantic knowledge it contains be exploited for document categorization. The traditional Bag of Words model (BOW) has several limitations that can be overcome by including external knowledge to the BOW document representation. Our aim is to design a *concept base generator* that can automatically augment short documents with a set of related concepts from Wikipedia.

Concept Base Generator

We propose two design approaches for developing concept base generator. The first design is based on a system for automatic linking of documents to encyclopedic knowledge indroduced by Mihalcea and Csomai (2007). The second design is based on a numerical statistic tf-idf.

Although Mihalcea and Csomai did not specifically work on document categorization, ideas for automatic keyword extraction and word sense disambiguation can very well be applied on this task. The approach based on their work is as follows: indentifying candidate phrases in the document, mapping them to surface forms of Wikipedia interlinks and disambiguating ambiguous phrases. The first step is to extract surface forms with corresponding articles. Resulting vocabulary will contain valuable phrases to be used for linking to the concepts. Given an input document, we find all such n-grams in it that appear in our vocabulary. Ambiguous phrases are further disambiguated using contextual overlap between the context of a phrase and candidate Wikipedia articles. The output is a set of related concepts (Wikipedia article titles) i.e.

concept base. We believe that augmenting a BOW representation with such concept base can considerably improve short-text categorization and greatly decrease error rates.

The second approach uses statistical variables term frequency (TF), document frequency (DF) and inverse document frequency (IDF). We propose to enumerate document frequency for each term in a dictionary, based on the number of occurences in Wikipedia articles. Each term in the dictionary should as well have a reference to articles it occurs in. Based on Wikipedia document frequency, we suggest to calculate TF*IDF for each term in the processed document. We would then take up to 20 terms with largest TF*IDF values and collect the articles they appear in. Concept base would then consist of titles of articles containing at least 25% of terms from the original document.

Implementation Details

For our experiment needs, we have parsed the Wikipedia XML dump, version of December 2011. We have removed topics with very short articles or no articles at all, events and disambiguation pages. Remaining 760,822 articles were left for concept generation. We processed the text creating a vocabulary of concepts which contained surface forms of all Wikipedia interlinks connected to each concept. We have also created a table of document frequencies for each unigram in the collection.

The keyword extraction algorithm was implemented to work in three steps: creating all n-grams from the input text (up to 4-grams); cross checking n-grams with the vocabulary to find candidate concepts; pruning candidates by removing the ones that already appear in the input text. Ambiguous candidates were further disambiguated using knowledge-based algorithm which relies on enumerating the contextual overlap between the concept and the ambiguous term. The

resultant set of concepts is augmented to the bag of words document representation as a concept base.

Evaluation

The proposed methods surpass the limitations of BOW model applied on short texts. Concept base generator is capable of enriching short documents with additional content-related concepts which leads to improvements in categorization performance. However, these improvements will be obtained at higher computational cost. Additionally, time constraints restricted the depth of this research which could include more scrupulous and empirical evaluation of proposed methods.

CHAPTER 05: RESULT AND FINDINGS

Wikipedia

Wikipedia is an free encyclopedia and polyglot of the Wikimedia Foundation (a nonprofit organization). Over 20 million articles in 282 languages and dialects have been drafted jointly by volunteers from around the world, and virtually anyone with access to the project can be an editor. Launched in January 2001 by Jimmy Wales and Larry Sanger, is the largest and most popular online reference work. Since its inception, Wikipedia has not only gained in popularity, it is among the 10 websites most popular in the world - but its success has given rise to sister projects. Among them, some have been accused of systemic bias and inconsistencies, with criticism focused on what some, like himself Larry Sanger , have agreed to call "anti-elitism" and that there is nothing but Project policy encyclopedic favor the consensus on the credentials in the editorial process . Other criticisms have focused on their susceptibility to being vandalized and the appearance of spurious information or lack of verification, although studies suggest that vandalism is generally disposed promptly.

There is also controversy over its reliability and accuracy. In this sense, the scientific journal Nature said in December 2005 that the English Wikipedia was nearly as accurate in scientific articles as the Encyclopedia Britannica . On the other hand and as stated in a report published in June 2009 by the Spanish newspaper El Pais , a study conducted in 2007 by Pierre Assouline, a French journalist, and conducted by a group of students of the Master of Journalism from the Institute of Political Studies in Paris analyzing the reliability of the project came in a book called The Wikipedia Revolution (Alliance), whose findings were quite critical. Among other things, stated that the Nature study was weak and biased, and that in his own study, the

Britannica was still 24% more reliable than Wikipedia. Of the 285 editions , sixteen exceed the 300,000 copies. The German version has been distributed on DVD-ROM , and intends to make an English version on DVD with over 2000 articles. Many of the issues have been replicated via Internet (using 'mirrors') and encyclopedias have given rise to derivatives (bifurcations) in other websites .

Features

The company culture has varied by state, in each version. In the main event, the Spanish Wikipedia any person has the ability to create a new item and almost any visitor can edit the content, except for items that are protected. However, in the English non-registered users can not start from scratch items. Wikipedia was created with the idea of producing quality text from the collaboration between users, like the development projects of free applications. Items evolve over time, and this is visible in its edit history. Usually, a portion of the edits are vandalism-content unrelated to Wikipedia or false information, and publishers sometimes have opposing views producing what is called edit war . This occurs when two or more publishers go into a cycle of mutual reversals due to disputes caused by differences of opinion on the content of the article. Do not confuse vandalism (which often affects one time to an item or items) to edit war, which affects repeatedly to the same item in a short time. Among the items vandalized frequently in the Spanish edition include: George W. Bush, Benedict XVI or Jehovah's Witnesses, while items with strong edit wars are Cuba or Valencia, because of the disparity between the views of its editors. Each chapter of Wikipedia has a group of staff, responsible for cooperation. Within this list are mentioned managers, whose main functions are to maintain, such as deleting items,

block vandals and other functions, and serve the fulfillment of the rules that govern it. The language version with the most administrators is the English Wikipedia with a total of over 1600.

Updating of information

Virtually all visitors can edit the content of Wikipedia and create new articles, and the changes are visible immediately after pressing the Save and approved by an authorized user - editor. The last condition was introduced in November 2008 to protect readers against the effects of vandalism. Wikipedia is built on the belief that cooperation between the users will lead to continuous improvement in the substantive content of passwords, in the manner in which this has been achieved in many open-source projects. Some editors of Wikipedia articles have described the editing process as an evolutionary process of social Darwinism.

Many people use the editing of Wikipedia to change the topic nonsense or eliminate vandalisms. However, since each edition is recorded in the history of the topic, all attempts to destroy can be detected and remedied. Model real-time collaboration enables editors to quickly complement existing topics. Sometimes, however, other ideas about the content of topics lead to so-called "edit wars", ie a state in which the editors of articles vary according to their own information, pulling each other's changes.

Wikipedia can be edited at any time. The ability to edit, however, may be temporarily blocked (or reduced, for example, only for registered users) due to frequent wars, vandalisms or editing. Wikipedia does not declare that any topic is "complete" or "finished", although it is planned to create a system of so-called stable version. The authors do not necessarily have any formal qualifications in the field to edit articles, also are informed that their contribution can be "freely edited and distributed" by everyone.

Active editors often create their own "watch list" entries on topics of interest to them in order to keep track of changes, including additions, discussions or potential vandalisms. Also, most previous versions can be read (and played), as is recorded in the "edit history" of the topic. This allows you to watch any version of the article, and their comparisons. The only exceptions are articles that have been removed. Edits concerning them are visible only to administrators of Wikipedia.

Computational linguistics

Computational linguistics is an interdisciplinary field of linguistics and computer science that uses computers to study and treat the language. To achieve this, attempts to model logically the natural language from a computational point of view. Such modeling does not focus on any area of linguistics in particular, but is an interdisciplinary field,
involving linguists , computer specialists in artificial intelligence , cognitive psychologists and experts in logic , among others.

Some of the areas of computational linguistics study are:

1. Corpus linguistic assisted by computer .
2. Design of parsers for natural languages .
3. Design of taggers or stemmers , such as POS-tagger.
4. Definition of specialized logic that serve as source for natural language processing .
5. Study the possible relationship between formal and natural languages.
6. Machine translation .

Computational linguistics (CL) or linguistic data processing (LD) examines how natural language in the form of text or voice data using thecomputer algorithmically can be processed. It is part of the field of artificial intelligence and the same interface
between linguistics and computer science .Theoretically, it is about clarifying the interplay or language and automation.

Like tf-idf to choose keywords to classify documents

The TF-IDF (Term Frequency - Inverse Document Frequency) is a weighting method often used in information retrieval , particularly in text mining . This statistical measure used to evaluate the importance of a term in a document in respect to a collection or corpus . The weight increases with the number of occurrences of the word in the document. It also varies depending on the frequency of the word in the corpus. Variants of the original formula are often used in search engines to assess the relevance of a document based on search criteria of a user.

Overview

The theoretical justification for this weighting scheme is based on empirical observation of the frequency of words in a text which is given by Zipf's Law . If a query contains the term T, a document is more likely to respond that it contains that word: the term frequency within the document (TF) is great. However, if the term T itself is very common in the corpus, that is to say that it is present in many documents (eg definite articles - the), it is actually little discriminating. Therefore the scheme proposes to increase the relevance of a term based on its rarity in the corpus (term frequency in the corpus low IDF). Thus, the presence of a rare term of the query in the document content is growing the "score" of the latter.

A naive Bayes model

Naive Bayesian model is a probabilistic method of instruction. The probability that a document d fall into the class c can be written as $P(c|d)$. Since the purpose of classification - to find the most appropriate class for this document, the naive Bayesian classification problem consists in finding the most probable class c m

$$c_m = \underset{c \in C}{\operatorname{argmax}} P(c|d)$$

Calculate the value of this probability directly is impossible because this requires a training set to contain all (or almost all) possible combinations of classes and instruments. However, using Bayes' formula, we can rewrite the expression for $P(c|d)$

$$c_m = \underset{c \in C}{\operatorname{argmax}} \frac{P(d|c)P(c)}{P(d)} = \underset{c \in C}{\operatorname{argmax}} P(d|c)P(c)$$

where the denominator is omitted, since it does not depend on c and, therefore, does not affect the determination of the maximum; P (c) - the probability that the class will meet c, regardless of the instrument in question; P (d | c) - the probability of finding a document d the class of documents c.

Using the training set, the probability P (c) can be estimated as

$$\hat{P}(c) = \frac{N_c}{N}$$

where N_c - The number of documents in class c, N - total number of documents in the training set. Here we use a different sign for the likelihood, as with the training set can only estimate the probability, but does not find its exact value.

In order to estimate the probability $P(d|c) = P(t_1, t_2, ..., t_{n_d}|c)$ Where t_k - A term from the document d, n_d - The total number of terms in a document (including repetition), it is necessary to introduce simplifying assumptions (1) of the conditional independence of the terms and (2) the independence of the positions of the terms. In other words, we neglect the first, the fact that the text in natural language the appearance of a word are often closely linked to the emergence of other words (for example, likely that the word integral encountered in the same text with a word equation than with the word bacteria) and, secondly, that the probability of finding one and the same word is different for different positions in the text. It is because of these gross simplifications, this model of natural language is called the naive (though it is quite effective in the classification problem). So, in light of the assumptions made, using the multiplication rule for probabilities of independent events, we can write

$$P(d|c) = P(t_1, t_2, ..., t_{n_d}|c) = P(t_1|c)P(t_2|c)...P(t_{n_d}|c) = \prod_{k=1}^{n_d} P(t_k|c)$$

Evaluation $P(t|c)$ using the training set will

$$\hat{P}(t|c) = \frac{T_{ct}}{T_c}$$

where T_{ct} - The number of occurrences of term t in all documents of class c (and in any position - there is essentially used the second simplifying assumption would otherwise have to calculate these probabilities for each position in the document that it is impossible to do accurately because of the sparsity of training data - it is difficult to expect that each term is met at each position enough times); T_c - The total number of terms in documents of class c. This figure includes all the re-entry.

Once the classifier is "trained", that is, the value found $\hat{P}(c)$ and $\hat{P}(t|c)$ you can find the document class

$$c_m = \underset{c \in C}{\operatorname{argmax}} \hat{P}(d|c)\hat{P}(c) = \underset{c \in C}{\operatorname{argmax}} \hat{P}(c) \prod_{k=1}^{n_d} \hat{P}(t_k|c)$$

To avoid overflow in the last formula below due to the large number of factors, in practice, instead of the product normally used sum of logarithms. Logarithm does not affect the determination of the maximum, since the logarithm is a monotonically increasing function. Therefore, in most implementations, instead of the last formula is used

$$c_m = \underset{c \in C}{\operatorname{argmax}} [\log \hat{P}(c) + \sum_{k=1}^{n_d} \log \hat{P}(t_k|c)]$$

This formula has a simple interpretation. Chances of a document classified as frequent a class above, and the term $\log \hat{P}(c)$ contributes to the total amount of a contribution. The values of $\log \hat{P}(t|c)$ the greater the more important the term t for the identification of a class c, and, accordingly, the greater their contribution to the total amount.

Summary of the existing research based on the semiotics framework

Area		Classification	Existing Research
Application domain		Domain specific	(Lanter 1991; Lanter and Essinger 1991; Alonso and Hagen 1997) in GIS, (Myers, Chappell et al. 2003b) in chemistry, (Bose 2002) in earth science, (Greenwood, Goble et al. 2003) in biology, (Cavanaugh, Graham et al. 2002) in physics, (Cui, Widom et al. 2000) in business intelligence.
		Domain independent	(Foster, Voeckler et al. 2002), (Szomszor and Moreau 2003), (Groth, Luck et al. 2004; Groth, Miles et al. 2005).
Data processing architecture		Database & file system	(Wang and Madnick 1990), (Woodruff and M. Stonebraker 1997), (Cui, Widom et al. 2000), (Buneman, Khanna et al. 2001), (Widom 2005), (Buneman, Chapman et al. 2006), (Muniswamy-Reddy, Holland et al. 2006).
		Service-oriented	(Foster, Voeckler et al. 2002), (Groth, Luck et al. 2004) , ((Greenwood, Goble et al. 2003; Zhao, Goble et al. 2003).
		Environment	(Myers, Chappell et al. 2003b), (Bose 2002).
Subject of provenance		Data-oriented	(Lanter 1991), (Buneman, Khanna et al. 2001), (Cui and Widom 2003), (Pancerella 2003), (Ram and Liu 2007).
		Process-oriented	(Foster, Voeckler et al. 2002), (Greenwood, Goble et al. 2003, (Zhao, Goble et al. 2003), (Frew and Bose 2001).
Focus area	Contents/semantics of provenance	Sparse	(Buneman, Khanna et al. 2001); (Cui and Widom 2003; Buneman 2006)
		Rich	(Myers, Pancerella et al. 2003b), (Zhao, Goble et al. 2003), (Groth, Luck et al. 2004).
	Harvest of provenance	Observation-based	(Reich, Liefeld et al. 2006), (Muniswamy-Reddy, Holland et al. 2006), (Buneman, Chapman et al. 2006)

Semantic similarity

Semantic similarity and semantic relatedness are often used synonymously. Both similarity and relatedness are measures of how close two concepts are to one another. Natural language processing (NLP) takes advantage of the measures for many applications, including

information extraction and retrieval, word sense disambiguation, text summarization, and type classification. However, these applications typically use similarity and relatedness interchangeably, which has led current research to focus mainly on semantic relatedness, when semantic similarity may be better than semantic relatedness or vice versa, as they are different. Specifically, semantic similarity is a subset of semantic relatedness. Similarity includes hyponymic and hypernymic relationships (is-a), while relatedness includes any and all functional relationships (has-a, is-a-part-of, etc.)[12]. The differences lead to several observations when determining a semantic similarity or semantic relatedness measure for a pair of concepts. Figure 1.1 illustrates the observations on how a concept pair falls into one of four areas with respect to the pair's measure of relatedness and similarity. It is not possible for a concept pair to fall into. Since similarity is a subset of relatedness, it is impossible to have a high similarity measure and a low relatedness measure. Where similarity and relatedness measures are both high, concept pairs like "George Washington" / "Abraham Lincoln" or "tiger" / "jaguar" fall firmly into area (2). The opposite holds for (3), where similarity and relatedness measures should both be low, like in the concept pair "nirvana" / "cheese grater". Both these areas give credence to the idea that similarity and relatedness are synonymous, as the measures are roughly equivalent with each other.

Wikipedia as Taxonomy

Wikipedia, in contrast, contains over 3.4 million articles in the English version alone as of November 2010. Wikipedia, like WordNet and other taxonomies, also contains internal structure and classification. These structures include the categories that an article belongs to as well as any internal links to other Wikipedia articles that are in the article text. Being open for

editing by anyone, Wikipedia grows incredibly quickly, and contains both esoteric and recent articles and concepts. Being a collaborative and open effort always means that Wikipedia has up to date information, but this means that the structure is not as well defined as a controlled taxonomy. This ill defined structure introduces variability that needs to be accounted for. The variability could be as simple as a mistake in classifying an article into a category, or may be as severe as deliberate vandalism, and in either case can affect the similarity or relatedness measure.

Mapping of Concepts

In all of the methods examined in this thesis document, Wikipedia is used as the taxonomy for determining the semantic similarity or semantic relatedness measurement. Since in NLP applications, the measurement is between two concept pairs, and Wikipedia consists of articles, the concepts must be mapped to Wikipedia. This usually means that the concept is equated with a single article from Wikipedia that best matches the context between the concept in the pair. Some methods map the concepts automatically, and some use a manual mapping. Regardless of whether it is done manually or automatically, there are three types of mapping that take place. First is a single direct correspondence, where a concept directly corresponds to an article in Wikipedia. An example of this mapping would be to map the concept “car” to Wikipedia. Wikipedia does not have an article named “Car”, but it does have an “Automobile” article that the concept directly corresponds to. Keeping this in mind we fail to reject H1 = Document classification can be improved using Wikipedia, H3 = Algorithm models are the factors that contribute to document classification, and H6 = the current position document classification is good.

CHAPTER 06: LIMITATIONS

One must be careful in answering the question of the research given the fact that first, due to the availability of limited literature on the subject area. Second, the case studies only represent responses from a number of people who may not have sufficient, reliable or authentic information. Not only the limited information obtained from both sources makes it very difficult for the research but makes the reader aware that the research tries to draw the most authentic and relevant conclusion to the problem. Additionally, time constraints restricted the scope and depth of this research which could include many more variables and elicit information from a larger population to be able to come up with a much stronger finding.

CHAPTER 07: CONCLUSION

Much existing literature for named entities focuses on semantic similarity between named entities as a method to enhance the identification of named entities from text. This task is called Named Entity Recognition (NER) and was formally described in the 6th Message Understanding Conference in 1995. One of the original types of named entities in NER is person names. To identify and disambiguate person names, Bunescu and Pasca utilized Wikipedia's article text and the categories that articles belonged to in order to drive a support vector machine (SVM). The taxonomy based kernel for the SVM took as an input a concept, then looked at that concept's text to identify and disambiguate to a set of possible articles. Finally it identified context based information contained in the possible articles' text and categories that matched the ambiguous concept. This method of using the category hierarchy and article text to find the similarity measurement between the concept and possible articles achieved an 84% disambiguation accuracy when applied to a disambiguation task. Cucerzan also used Wikipedia article text and category pages to create a system that used semantic similarity for entity identification and disambiguation. The method employed by Cucerzan was to process the contextual information contained in the concept, then match the context of candidate articles and their category information in order to maximize the agreement between the article context and the concept context. The most similar article to the concept context was selected as the disambiguated article. Cucerzan obtained accuracy results of 88% to 91% on disambiguation tasks dealing with named entities.

Cosine similarity is used in many applications, perhaps most often in text mining. It works well to find the similarity between TF-IDF vectors. We selected cosine similarity because

it was used in other similarity and relatedness applications, both in those that use Wikipedia like ESA and WLM as well as those that use more traditional taxonomies. The cosine similarity based algorithm in Figure 3.9 below was the most computationally complex algorithm, as the dot product and Euclidean distance for both category vectors needs to be calculated. This creates a much slower overall execution speed than either the Dice or Jaccard methods. Notice, like the intersection vectors from the Dice and Jaccard algorithms, the individual category vectors are extended out to include the categories from the other category vector. The new categories in the category vectors have a value of 0, and do not affect the dot product or euclidean distance. Interestingly, due to the extension of either a category vector or the creation of an intersection vector, all three algorithms start out with the same steps, creating a union vector. This is probably not strictly necessary and definitely hurts the execution speed of both Dice's coefficient and the cosine similarity, as neither algorithm needs the union.

We proposed a way to enrich the BOW representation of processed documents using the semantic knowledge from Wikipedia. Although the proposed methods need deeper evaluation, they seem to surpress short-text documents limitations that BOW model has.

Wikipedia database

What is missing an examination of how SNA can provide insight into collaborative knowledge construction when the authors are largely unknown and inaccessible? In particular, using an SNA approach to provide a method by which we can visualize the contributions of unknown authors and draw conclusions about motivations, their level of content knowledge, the authority of the information they share, and the overall legitimacy of the content. In other words, we may be able develop an approach using SNA that can be used to generate profiles of

unknown contributors to a collaborative project and use those profiles to inform us regarding the legitimacy of the content. In Wikipedia the anonymous authors of articles could be considered from an SNA perspective. The relationships between authors and articles form their own nodes and distances.

An article on Quantum Mechanics, for example, is spatially closer to an article on Einstein's Theory of Relativity, by virtue of being from a related field of science, than it is to an article on Biology, a different area of science, or an article on Einstein himself (biography) or any other non-science article. Authors can also be spatially related to other authors and articles. Authors of the same article are very closely related. We can reasonably expect that many authors will contribute to more than one article and that their contributions are also spatially related (Korfiatis, Poulos, & Bokos, 2006). An author who contributes regularly to an article on Quantum Mechanics may have a strong background in science and may contribute to other articles related to science. Contributions to similar articles would be closer together spatially than article contributions in disparate fields. An author's pattern of contributions and their spatial relatedness may be used to make inferences about an author's level of knowledge and, by extension, the overall authority and legitimacy of an article. In studies using SNA, the actors are generally known or knowable to some degree or, in other words, researchers usually have access to actors and are able to question them directly or indirectly. What has been less studied is networks and relationships between actors who are largely anonymous and known only by pseudonyms and indirectly through the information they exchange.

Application of document classification:

1. spam filtering
2. compilation of online catalogs
3. selection of content
4. systems documentation
5. automatic abstracting (drawing annotations)
6. removing the ambiguity in the automatic translation of texts
7. limiting the search in the search engines
8. definition of encoding and language of the text

References

A. Huang, D. Milne, E. Frank and I. H. Witten, "Clustering Documents using a Wikipedia-based Concept Representation," Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, pp. 839-844, 2008.

B. Mann, "Annotation of special structures in astronomy," presented at Workshop on Data Derivation and Provenance, Chicago, Illinois, 2002.

C. Goble, "Position statement: Musings on provenance, workflow and annotations for bioinformatics," presented at Data provenance/derivation workshop, 2002.

C. Pancerella, "Metadata in the Collaboratory for Multi-scale Chemical Science," presented at DC-2003: the 2003 Dublin Core Conference, Seattle, Washington, 2003.

D. Ballou, R. Y. Wang, H. Pazer, and G. K. Tayi, "Modeling information manufacturing systems to determine information product quality," Management Science, vol. 44, pp. 462-84, 1998.

D. Lanter and R. Essinger, "User-centered graphical user interface design for GIS," National Center for Geographic Information and Analysis, UCSB 91-6, 1991.

D. Lanter, "Design of A Lineage-Based Meta-Data Base For GIS," Cartography and Geographic Information Systems, vol. 18, pp. 255-261, 1991.

D. Pearson, "The Grid: Requirements for Establishing the Provenance of Derived Data," presented at Workshop on Data Derivation and Provenance, Chicago, Illinois, 2002.

E. Gabrilovich and S. Markovitch, "Overcoming the Brittleness Bottleneck using Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge," Proceedings of the Twenty-First National Conference on Artificial Intelligence, pp. 1301-1306, 2006.

G. Alonso and A. El Abbadi, "Goose: Geographic object oriented support environment," presented at The ACM Workshop on Advances in Geographic Information Systems, Arlington, Virginia, 1993.

G. Alonso and C. Hagen, "Geo-opera: Workflow concepts for spatial processes," presented at 5th International Symposium on Spatial Databases, Berlin, Germany, 1997.

J. Frew and R. Bose, "Earth System Science Workbench: A Data Management Infrastructure for Earth Science Products," presented at the 13th International Conference on Scientific and Statistical Database Management, Fairfax, VA, 2001.

J. Frew and R. Bose, "Lineage Issues for Scientific Data and Information," presented at Data provenance/derivation workshop, 2002.

J. L. Romeu, "Data Quality and Pedigree," AMPTIAC 1999.

J. Myers, A. Chappell, M. Elder, A. Geist, and J. Schwidder, "Re-integrating the Research Record," IEEE Computing in Science & Engineering, vol. 5, pp. 44-50, 2003.

J. Zhao, C. Goble, M. Greenwood, C. Wroe, and R. Stevens, "Annotating, Linking and Browsing Provenance Logs for e-Science," presented at 2nd Intl Semantic Web Conference (ISWC2003) Workshop on Retrieval of Scientific Data, Sanibel Island, FL, 2003.

K. Stamm and R. Dube, "The Relationship of Attitudinal Components to Trust in Media," Communication Research, vol. 21, pp. 105-123, 1994.

L. Ding, P. Kolari, T. Finin, A. Joshi, Y. Peng, and Y. Yesha, "On Homeland Security and the Semantic Web: a Provenance and Trust Aware Inference Framework," presented at AAAI Spring Symposium on AI Technologies for Homeland Security, Stanford University, CA, 2005.

M. Ceruti, S. Das, A. Ashenfelter, G. Raven, R. Brooks, M. Sudit, G. Chen, and E. Wright, "Pedigree Information for Enhanced Situation and Treat Assessment," presented at the 9th International Conference on Information Fusion (ICIF 2006), Florence, Italy, 2006.

M. Janik and K. Kochut, "Training-less Ontology-based Text Categorization," presented at the Workshop on Exploiting Semantic Annotations in Information Retrieval at the 30th European Conference on Information Retrieval, Glasgow, Scotland, 2008.

M. Janik and K. Kochut "Wikipedia in Action: Ontological Knowledge in Text Categorization," presented at the 2008 IEEE International Conference on In Semantic Computing, pp. 268-275, Santa Clara, USA, 2008.

M. Greenwood, C. Goble, R. Stevens, J. Zhao, M. Addis, D. Marvin, L. Morea u, and T. Oinn, "Provenance of e-Science Experiments - Experience from Bioinformatics," presented at UK e-Science All Hands Meeting, Nottingham, UK, 2003.

N. Prat and S. Madnick, "Evaluating and Aggregating Data Believability across Quality Sub-Dimensions and Data Lineage," presented at WITS 2007, Montreal, Canada, 2007.

P. Buneman, S. Khanna, and C. T. Wang, "Why and Where: A Characterization of Data Provenance," in Lecture Notes in Computer Science, vol. 1973, V. V. Jan Van den Bussche, Ed.: Springer, 2001.

P. Buneman, S. Khanna, and W. C. Tan, "Data Provenance: Some Basic Issues," presented at FSTTCS, New Delhi, India, 2000.

P. Wang and C. Domeniconi, "Building Semantic Kernels for Text Classification using Wikipedia," Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 713-721, Las Vegas, USA, 2008.

R. Bose, "A Conceptual Framework for Composing and Managing Scientific Data Lineage," presented at 14th International Conference on Scientific and Statistical Database Management, 2002.

R. Cavanaugh, G. Graham, and M. Wilde, "Satisfying the Tax Collector: Using Data Provenance as a way to audit data analyses in High Energy Physics," presented at Workshop on Data Derivation and Provenance, 2002.

R. Mihalcea and A. Csomai, "Wikify!: Linking Documents to Encyclopedic Knowledge," CIKM '07 Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, pp. 233-242, 2007.

S. Ram and J. Liu, "W7 Model: an Ontological Model for Capturing Data Provenance Semantics," in Lecture Notes in Computer Science 4512, L. Wang and P. Chen, Eds.: Springer, 2007.

W. Tan, "Research Problems in Data Provenance," IEEE Data Engineering Bulletin, vol. 27, pp. 45-52, 2004.

Y. Cui and J. Widom, "Lineage tracing for general data warehouse transformation," VLDB Journal, vol. 12, pp. 41-58, 2003.

Y. Cui, J. Widom, and J. Wiener, "Tracing the Lineage of View Data in a Warehousing Environment," ACM Trans. Database Syst., vol. 25, pp. 179-227, 2000.

Y. Simmhan, B. Plale, and D. Gannon, "A Survey of Data Provenance Techniques," Technical Report IUB-CS-TR618, Indiana University, 2005.

Y. Simmhan, B. Plale, and D. Gannon, "A Survey of Data Provenance in e-Science", ACM SIGMOD Record, vol. 34(3), September 2005