THE IMPACT OF WIKIPEDIA ON SCIENTIFIC RESEARCH

Richard Khoury

Department of Software Engineering, Lakehead University, Thunder Bay, Canada
richard.khoury@lakeheadu.ca

ABSTRACT

Recently, many researchers have conducted scientific studies with, about, or using the online encyclopaedia known as Wikipedia. While they have obtained promising results and reached interesting conclusions, their work has been limited to their own specific fields. This paper presents a survey of the emerging branch of Wikipedia-based research. Our aim is to put these studies into a broader scientific context, to show the extent of the work done, its limitations, and suggest some future directions.

KEYWORDS

Data-mining, e-Society, Internet applications, Text Processing, Wikipedia

1. INTRODUCTION

Since 2002, a number of researchers have independently started using Wikipedia as a knowledge source for applications in a large range of projects, from natural language processing (NLP) [1] to social studies [2], and the interesting results they published in over 300 papers and book chapters underlie the potential of Wikipedia. To paint a broad picture of Wikipedia's scientific impact, this paper presents a wide-ranging survey of the emerging branch of Wikipedia-based research. In line with this objective, Section 2 provides an overview of representative projects that use Wikipedia in different fields of study. We selected these papers to illustrate the diversity of projects that can benefit from Wikipedia, the various ways Wikipedia can be adapted to serve certain functions, and when applicable how the results obtained compare with those of state-of-the-art systems. In Section 3 we present the main arguments cited by various authors to justify the use of Wikipedia in academia; in Section 4, as a counter-point, we present the main flaws that researchers discovered when studying that resource. In Section 5, we discuss how Wikipedia can further contribute to scientific research, and finish with some concluding remarks in Section 6.

2. SCIENTIFIC IMPACT

2.1. In Natural Language Processing

Projects in NLP are among those that have benefited the most from the rise of Wikipedia. In a first example of such projects, Schönhofen [1] proposed the use of Wikipedia's category hierarchy as class labels for document classification. In Wikipedia, categories represent nearly 200,000 semantically-meaningful topic groups [3] of varying granularity that are covered by each article. Consequently, Schönhofen's system begins by discovering all articles possibly relevant to the document to classify, and then weights the resulting list of categories. To summarize the results presented in [1], the system classifies 88% of the 20 Newsgroup posts and 70% of RCV1 news articles into the correct Wikipedia categories. Another common challenge in NLP applications is that of computing the similarity between words encountered in text documents. Towards that end, Wee and Hassan [4] developed a simple but effective way of computing the directional similarity between two words, based on the ratio of the number of

Wikipedia articles containing both words to the total number of articles in which one of them appears. Their results show a 9% improvement in accuracy over the next-best algorithm in the literature. They credit this gain in part to the fact that, thanks to Wikipedia's sheer size, their algorithm could compute the similarity between many more pairs of words than other systems.

In an example of a more complex NLP task, Ahn *et al.* [5] used Wikipedia in an automated question-answering (QA) system. Their system begins by finding a relevant Wikipedia article for the question. It then scans the article for named entities and ranks them, and returns the named entity with the highest score as the answer. The authors' preliminary results show a promising 14% increase in F-measure compared to the TREC 2004 median. Another well-known NLP challenge is that of named entity disambiguation (NED). This is typically accomplished by collecting text documents in which the entities are labelled unambiguously and using them to discover the "context", a window of words surrounding each name. Then, when an ambiguous name is encountered, the similarity between the surrounding words and the different known contexts is used to resolve the ambiguity. Bunescu and Pasca [6] enhanced this technique by using Wikipedia's titles to list proper names and the related article's text to gain context. This generated a dataset of over 1.7 million disambiguated named entities which could be used in a classic NED system. They then enriched their dataset by using the Wikipedia category structure. Bunescu and Pasca found that using this extra information gives a 16% improvement in disambiguation accuracy.

Wikipedia is also suited to multilingual NLP. Indeed, versions of the encyclopaedia are growing in 248 different languages [7]. Moreover, some articles covering the same subject in different languages are connected by *cross-language links*, links prefixed by the target language's ISO639 code. This turns Wikipedia into an aligned corpus of several hundred languages, a rare resource in NLP which many researchers have exploited. For example, Sorg and Cimiano [8] used them to map queries between English, German and French resources. They experimented with a set of aligned trilingual queries, and used the tf.idf value to match each query to the best cross-linked articles in its language. They then followed the links to retrieve the matching articles and compared the ranks in the other languages. Although the results were mixed – some article pairs had similar ranks while others' were wildly different – the authors conclude that there is a lot of potential for future research.

2.2. In Knowledge Acquisition

Though it often goes hand-in-hand with NLP, research in knowledge acquisition focuses on the challenge of mining semantic information from input sources such as text corpora, databases and humans, and representing it in some structured form, such as a taxonomy [3] or a thesaurus [9]. Wikipedia's category hierarchy gives an interesting starting point for such a taxonomy, which can then be enriched in a number of ways. For example, Nastase and Strube [3] noted that certain category names are short phrases, which can be parsed in order to discover the semantic relationships, such as the *isa* and *member_of* relations, that exist between the various concepts in the taxonomy. This allows them to add meaningful semantic links to their taxonomy. Their experiments found that, up to 98% of the time, these links match those a human annotator would have picked. Another alternative is to use Wikipedia to enrich another existing taxonomy, such as WordNet. Wu and Weld [10] did this by exploiting infoboxes, templates where key information on common topics is labelled and summarized. After cleaning up the information, they map the infoboxes to the WordNet nodes with the most similar names. The resulting taxonomy combines the information found in Wikipedia, such as the birthplace of performers, with the reliable structure of WordNet, which includes for example the fact that "performer" is synonymous with "performing artist" and that "birthplace" is a location, and can thus infer more complete answers to queries like "which performing artists were born in Chicago?"

The structure of Wikipedia also bears similarity to that of traditional thesauri. Specifically, the synonymy, hierarchy and associative relation between words in a thesaurus exist in Wikipedia in the form of redirect pages, the graph of categories, and the links between different articles respectively. By exploiting these structural similarities, Milne *et al.* [9] were able to automatically build a thesaurus. When they compared it to Agrovoc, a manually-created thesaurus for the food and agriculture domain, they found that their thesaurus only covered the 50% most commonly-used terms found in Agrovoc. However, when the researchers extracted noun phrases from a corpus of agriculture-related documents and searched for them in Agrovoc and in their Wikipedia-based thesaurus, they found three times as many terms in their thesaurus as in Agrovoc. This result is due to the fact that many of the terms encountered in the documents fall outside Agrovoc's domain, but are part of Wikipedia and therefore part of their thesaurus. This leads the authors to judge that their thesaurus outperforms Agrovoc [9]. Wang *et al.* [11] further used the thesaurus to enrich text documents prior to classification, by adding relevant concepts present in the thesaurus' entry but missing from the document Their experiments show that use of the enriched document text yields a 3% to 7% improvement in the classification results compared to using the original text alone.

2.3. In Social Studies

A growing body of research focuses on studying the various aspects of social interactions within the Wikipedia community. In their research, Brandes and Lerner [2] examined the development of antagonistic relationships between users. Thanks to their graphical revision networks, they were able to visually observe patterns in controversial articles' edit history, such as the different sides of the gun-rights debate fighting over the "Gun politics" article. A similar graphical approach was also advocated by Suh *et al.* [12]. They catalogued different behaviours, such as clusters of Wikipedia editors (called *Wikipedians*) forming on different sides of a controversial article, and these clusters spreading to other, related articles. Viégas *et al.* [13] proposed their own graphical representation, which tracks the editing of a page by different contributors over time. They were thus able to discover some interesting patterns, such as the tendency for the initial text of an article to lasts longer and receive fewer edits than other contributions.

Viégas *et al.* [14] studied the social impact of the talk pages associated with each article. They found that these special pages served several important functions within the Wikipedian community. First and foremost, these pages are used to plan and discuss the evolution of the related articles and thus create some much-needed behind-the-scenes coordination in the project. But moreover, participating in this discussion forum strengthens the sense of community within Wikipedia. Finally, as Wikipedians often quote Wikipedia policies relevant to the discussion, talk pages serve to disseminate knowledge about Wikipedia guidelines through the community.

2.4. Research about Wikipedia

Some researchers go beyond using Wikipedia as a language or social resource, and actually consider it a topic worthy of study in and of itself. For example, Thomas and Sheth [15] studied the changes over time in Wikipedia's more mature *good articles*. They adopted Wikipedia's criterion of maturity, which is stability, or that no major changes have occurred on the page despite a large number of edits. Their study found that mature articles really had converged to a stable state over time, and that a similar pattern of convergence exists in regular articles, which makes it possible to evaluate the current stability and maturity of any article. Alternatively, Lim *et al.* [16] developed a peer-review metric to evaluate the quality of articles. They begin by defining an editor's authority as proportional to the amount of content he edited in various articles and which is still part of the current version of the articles. They then computed the article's peer-reviewed score as a function of the proportion of its content originating from authoritative editors.

Viégas [17] studied the related topic of images used in Wikipedia articles. These are not strictly part of Wikipedia, as the wiki software is limited to text editing. Rather, users edit images using their own computers and software, and upload them to a database named WikiCommons. Viégas found that users were creating and donating professional-quality photographs and scientific images, often for no reason other than they felt an article would benefit from it. Moreover, despite being isolated from the main Wikipedia editing process by technical limitations, image editors still shared in the Wikipedian sense of community.

3. ARGUMENTS FOR USING WIKIPEDIA

3.1. Size

The most often-cited reason in the literature to use Wikipedia as a resource is its size and width of coverage [1], [4], [5], [7], [9], [11]. Many practical applications require access to a large and semi-structured knowledge base, such as a general or domain-specific encyclopaedia. But researchers prefer to use the largest semantic resource available, and in most cases that is Wikipedia. To illustrate, we present in Table 1 a comparison of the size of various popular encyclopaedias. It can be seen from these results that Wikipedia has a clear advantage over most other resources, with the exception of a few very specialized resources such as the Guide Star Catalog II. But save for such exceptional cases, Wikipedia remains a much larger resource for research in most other domains as well as for general, non-domain-specific applications.

Table 1. Comparison of the size of various encyclopaedias.

Encyclopaedia	**Description**	**Size**
Wikipedia English version, as of 15 May 2008	Online general encyclopaedia	2,373,734 articles
Encyclopædia Britannica 32-volume edition, 2007	Printed general encyclopaedia	Over 65,000 articles
Microsoft Encarta	Digital general encyclopaedia	Over 42,000 articles
The Great Soviet Encyclopedia Third edition, 30 volumes, 1969-1978	Printed general encyclopaedia	94,541 articles
Oxford English Dictionary Second edition, 20 volumes, 1989	Printed English dictionary	Over 600,000 words
American Medical Association Complete Medical Encyclopedia, 2006	Printed medical encyclopaedia	Over 5,000 articles
The Catholic Encyclopedia 16 volumes, 1907-1914	Printed religious encyclopaedia	Around 11,500 articles
Guide Star Catalog II Version 2.3.2, 2005	Digital astronomical catalogue	945,592,683 objects
MathWorld As of 15 May 2008	Online mathematical reference work	12,834 entries

We have mentioned in Section 2.1 that Wikipedia is also used as a corpus of text documents for NLP research. Once again, this is a field where using a larger corpus is preferable, as it will allow researchers to compute more accurate language statistics. In line with that observation, we present in Table 2 a comparison of the size of Wikipedia and of a few of the most popular corpora uses in NLP research today. In almost all cases, Wikipedia clearly has a massive size advantage over traditional NLP corpora. But as before, there are exceptions. In this case, we can note the TREC conference Web Track, which is necessarily large in order to evaluate web searching techniques.

Table 2. Comparison of the size of various NLP corpora.

Corpus	Size
Wikipedia, English version, 15 May 2008	2,373,734 articles, over 1 billion words
Brown Corpus, 1964	500 samples, over 1 million words
WordNet 3.0, 2006	117,659 synsets, 147,278 words
British National Corpus (BNC), 1994	4,054 texts, over 100 million words
SWITCHBOARD-1, 1993	2,430 conversations, about 3 million words
TREC Web Track, 2009	1 billion Web pages

3.2. Growth

Wikipedia is not only larger than most other encyclopaedias and corpora; it is also growing at a much faster rate. Indeed, NLP corpora are typically either static or grown by a small team of experts. From the examples in Table 2, we can note that no new texts have been added to the Brown Corpus or the BNC since their creation, while additions to WordNet are made by a team of researchers at Princeton University, while people who wish to contribute to SWITCHBOARD are required to satisfy a set of requirements and to contact the project directly. Meanwhile, updating an encyclopaedia is a time-consuming and expensive process, which typically involves domain experts, professional editors, and a lengthy review process. It is easy to see, then, how Wikipedia's open-door policy on creating and contributing to articles leads to a much greater growth rate.

Wikipedia's growth was the subject of two studies. Voss [7] studied a period going from Wikipedia's creation in 2001 to 2005. He found that, following an initial year of linear growth, the project started growing exponentially. This can be observed in all aspects of Wikipedia, including the number of articles, their size, and the number of active Wikipedians. However, this exponential growth is interrupted by brief periods of linear growth, which Voss explains as the project reaching its hardware and software limits, and slowing down until an upgrade can be made. This implies that, for technical reasons, Wikipedia's growth cannot be sustained indefinitely [7].

The research by Buriol *et al.* [18] covers a three-year period from 2003 to 2006. Like Voss, these authors observed an exponential growth in the number of articles and of Wikipedians. However, when they studied the articles by creation date, they found that older articles tend to undergo linear growth. The authors hold this to be a sign that Wikipedia's exponential growth is part of a transient growth phase, which cannot be sustained endlessly. Moreover, in light of their analysis and contrary to Voss, they believe that Wikipedia has already started showing signs of reaching its maturity phase [18].

It is interesting to note that, from January 2003 to January 2006, the number of articles in the English version of Wikipedia increased from 106,000 to 942,000, giving a growth rate of nearly 789% over three years. By contrast, WordNet 2.0, released in 2003, counted 115,424 synsets and 144,309 words. Comparing to the 2006 statistics of WordNet 3.0 presented previously gives only a 2% growth rate over the same period. This illustrates how impressive Wikipedia's growth rate is compared to that of another popular NLP resource.

3.3. Mark-up tags

One last advantage to using Wikipedia compared to another text corpus is that the articles' source code makes heavy use of mark-up tags, which makes them a lot easier for an automated system to handle [1], [6], [9], [11]. Unique tags mark the title of each subsection, as well as special content such as templates, categories, images, references, mathematical equations, and external links. Within the text, important concepts and named entities are *wikilinked*, which means that the keywords in the text are linked to the relevant articles. These links are clearly labelled in the source code using double square brackets. Wikilinks provide researchers with a growing source of information: the number of links in Wikipedia is increasing exponentially [7].

However, one cannot assume that the opening and closing tags will always be balanced in all articles. Given Wikipedia's open-door editing policy and the absence of any editorial oversight, errors can occur. Consequently, one must take steps to verify the tags for consistency before or during processing. Still, despite this minor setback, Wikipedia's tagged information has been successfully used in a number of applications. In our previous examples, Bunescu and Pasca [6] use wikilinks to build a dictionary of named entities, and the automated thesaurus-building projects we presented make use of both wikilinks and category links to detect different types of semantic relationships [9], [11].

4. Problems with using Wikipedia

4.1. Irregular coverage

Many developers make a point of balancing their NLP corpora by giving equal representation to all the domains composing it. Alternatively, some corpora, such as the Brown Corpus [19], provide a varying level of representation to each domain according to some carefully-defined proportional scheme. These representation schemes are useful when using the corpora to train NLP systems. By contrast, Wikipedia is not a balanced corpus, nor does it make any effort to become one.

The growth of Wikipedia's domain coverage can be understood from two perspectives. First, from an individual perspective, a Wikipedian will create and contribute to articles on topics he knows and has an interest in. Predictably enough for an online open-source online project, the average contributor tends to be scientific-minded and very tech-savvy. Consequently, the coverage of science and technology topics is much deeper than that of topics related to arts and the humanities [20]. A Wikipedian will also contribute information known from personal experience. Given that the average age of an editor is around 31 years old [7], this makes Wikipedia heavily biased towards recent events. This explains some of the oddly selective coverage observed in some studies. For example, a history professor who reviewed Wikipedia's entry on women's rights in the United States was surprised to find a discussion of Valerie Solanas but no mention of the 19th Amendment [21]. But we can note that Solanas died in 1988, was the subject of a movie in 1996, and her play posthumously debuted in 2000, while on the other hand the 19th Amendment was ratified in 1920 and has not been publicly debated or contested for generations.

Second, from a group perspective, articles on more popular topics are read more often by more people, and are therefore updated more frequently [18]. It follows that popular domains are more detailed in Wikipedia than more obscure or specialized domains [21]. Popular culture topics are among those that receive the most attention from the public. Articles in these domains tend to be very detailed and divided into a great number of fine-grained children articles. For example, for a given popular band, there will also be an article for each band member, each album, and each hit song. In contrast, less popular topics receive much less attention and are much less developed. Wikipedia is far from having one article for each university, much less one for each university professor; when it does, the article is typically very brief and lacking detail. To give an example of this irregularity, Greenstein [22] notes that the article for the fictional space explorer Jean-Luc Picard is much more detailed than the article for Patrick Stewart, the actor portraying him, and that Stewart's article is considerably longer than that of real-life space explorer John Huchra. It is worth updating this observation: one year and several hundreds of edits later, Stewart's article has surpassed Picard's in length and detail, thanks mostly to the enforcement of a policy to cut down on the un-encyclopaedic level of detail in the lives of fictional characters. Meanwhile, Huchra's three-line article was edited only five times in the same time span, and only for maintenance purposes.

4.2. Inaccurate information

The factual accuracy of Wikipedia's articles is possibly one of the most hotly debated issues surrounding the encyclopaedia. Defenders believe that, given enough editors, all mistakes will be found and corrected [7], [16]. A similar assumption was adopted by Lim *et al.* [16], when they measured the quality of an article in function of the number of authoritative editors who worked on it. However, while Lim *et al.* measured and ranked editors, Wikipedia considers all editors as equals, and welcomes everyone to edit articles without any qualification checks whatsoever. But saying that a large number of people worked on an article is meaningless when it is impossible to know if any of them were actually qualified to contribute to it [7].

In a famous incident illustrating this problem, for over four months the Wikipedia article of journalist John Seigenthaler Sr. included an allegation that he was directly involved in both Kennedy assassinations [20]. During these four months, no one who edited the article recognised that the libellous statement was made-up, until a friend of Seigenthaler pointed it out. This highlights the fact that the number of editors who go over an article is not a measure of accuracy. What matters is the number of qualified editors, and that is not a distinction Wikipedia makes. Even worse, established real-world authorities who have written articles in their area of expertise have seen their contribution treated equally to that of, and edited by, unqualified lay contributors [20]. These experts get discouraged by seeing their additions overlooked, cut down, or even "corrected" by unqualified editors, and tend to give up working on the project entirely [20], [23].

One cannot discuss the topic of Wikipedia's factual accuracy without mentioning the now-famous 2005 study published in Nature comparing Wikipedia to Encyclopaedia Britannica [24]. In their investigation, Nature selected 42 scientific entries from Wikipedia and Britannica and sent them to experts for review. The experiment revealed that Wikipedia came close to Britannica in terms of accuracy; Wikipedia showed an average of four inaccuracies per article, while Britannica had three. This surprising result has since been held by the Wikipedia community as proof that their populist way can almost equal the efforts of renowned experts in the world's most famous reference work. Britannica, on the other hand, contested the article in no uncertain terms, calling the study flawed and poorly carried out, and the results error-laden and meaningless [25]. Britannica accuses Nature of falsifying its data by editing its Encyclopaedia articles, cutting and merging several entries together and sometimes writing in new text, and by taking articles from other Britannica publications and falsely presenting them as Encyclopaedia entries [25]. Moreover, Nature's notion of an "inaccuracy" was left vague. Minor errors in Britannica, such as an ambiguity in the number of siblings of a scientist, were considered inaccuracies equal to critical scientific mistakes in Wikipedia. Nature accepted its reviewers' comments without double-checking and thus counted a number of errors against Britannica when in fact it was the reviewers who were mistaken. Omissions were also universally counted as inaccuracies against Britannica, even if they were subjective (such as when a reviewer felt that an extra equation could help illustrate an article), or part of a different article (such as mentioning Nobel Prize laureates in each individual's biography rather than the Nobel Prize article), or even the result of Nature's editing of the article (the 6000-word-long Britannica article on lipids was cut to 350 words by Nature, and the omissions noted by the reviewer and counted in Britannica's inaccuracy score are actually part of the missing 5650 words) [25]. Such a level of poor scientific conduct makes Nature's conclusions valueless. Despite this fact, they unfortunately continue to be cited as fact by many authors [2], [4], [14], [15], [16].

Rosenzweig [21] attempted a similar experiment using history articles. He studied 25 biographies of American historical figures, and found factual errors in only four of them, a result that puts Wikipedia on par with Microsoft's Encarta encyclopaedia. However, Rosenzweig points out that these results cannot be generalized to the whole of Wikipedia. He notes, as we did in the previous section, that Wikipedia articles are developed following popular interests, and that historical biographies are a popular topic. Other types of historical articles were far more incomplete and inaccurate; he found that the one on the history of US immigration, for example, "verges on incoherence" [21].

4.3. Vandalism

Vandalism is one of the most well-known and widely acknowledged problems facing Wikipedia today. In their research, Viégas *et al.* [13] observed and defined six common kinds of vandalism. The first and most obvious is mass deletion, or the complete deletion of an article's text. There are two additional variations: obscene mass deletion, when the content of an article is completely replaced by vulgarities, and phony redirection, when an article is replaced with a redirect to an unrelated article. The other three kinds of vandalism do not involve a mass deletion. Offensive copy refers to the addition of vulgarities within the text of an article, while phony copy refers to the addition of off-topic text. The last and most subtle kind of vandalism is idiosyncratic copy, which refers to the addition of text to an article that can be seen as relevant, but is also one-sided, inflammatory, or false.

Viégas *et al.* [13] focused in their research on the three variations of mass deletion. They found that the median time for detection and rectification of these acts of vandalism was less than three minutes, while the mean time for corrections was a little under a week. This means that some of mass deletions endure long enough to have a noticeable impact. Follow-up research by other authors [18] confirms that 70% of mass deletions were corrected in less than one hour. Both studies lead to the same conclusion: the vast majority of mass deletions are detected and corrected swiftly, but some outliers endure.

The other three kinds of vandalism have not yet been the topic of statistical analysis, as they require actually reading and understanding the content of the article. However, in a smaller-scale study reported in [20], Halavais experimented with idiosyncratic copy vandalism by deliberately inserting 13 errors in Wikipedia articles, ranging from the obvious to the obscure. To his surprise, all 13 claims were corrected in less than three hours. But once again, not all acts of idiosyncratic copy are corrected so swiftly. We have previously presented the case of John Seigenthaler Sr., whose similarly-vandalized article went uncorrected for over four months [20].

4.4. Lack of standard test results

As research using Wikipedia is still in its infancy, it can be difficult or impossible for authors to compare their results to literature standards. In [1], for example, the author classified the documents of two standard test corpora into Wikipedia categories. Unfortunately, all previous works used a completely different taxonomy, which makes direct comparisons impossible. On the other hand, [6] created their own benchmarks by applying the standard cosine method to Wikipedia articles to verify the benefit of their NED method using Wikipedia's category hierarchy. It is worth noting however that this is not a universal problem. For example, the results of [4], [5] and [9] are compared to the literature benchmarks. Moreover, as more research is done using Wikipedia, we can realistically expect more results to be published and new benchmarks to emerge naturally.

5. Future research directions

With over 300 publications in various fields of study over the past 6 years, a lot of groundwork has been done on Wikipedia-based research. Yet, in our opinion, this work only scratches the surface of Wikipedia's potential as a resource. Indeed, most projects exploit only a single semantic aspect of Wikipedia – the wikilinks, the categories, the infoboxes, and so on. Yet projects that combine several aspects lead to more complex applications. Contrast for example the wikilink-based NED system [6] or the category-based classifier [1] with the automated thesaurus-building system resulting from both wikilinks and categories [9]. On the other hand, while some authors have already begun enriching existing systems with Wikipedia [10], much work remains to be done. For example, automated translation systems could benefit from using cross-language wikilinks to recognise and handle complex scientific or technical expressions whose equivalents are not word-for-word translations. Finally, Wikipedia itself could benefit, for example from a project to automatically verify the information found in its articles and infoboxes against more established and reliable sources.

This field of research also leads to practical applications. Wissner-Gross [26] already proposed a tool to generate custom reading lists of relevant Wikipedia articles based on a user's topic of interest. However, given our discussion in Section 4.2 on Wikipedia's inaccuracy as well as the poor writing style resulting from its writing-by-committee approach [21], it could be preferable to build a reading list of more authoritative sources. This could be achieved by exploiting the reference citations included in most articles. References are tagged in the article's code, so they can easily be extracted, and they will often point to newspaper articles, books, and peer-reviewed publications. By using these references, one could create a reading list of much more authoritative, accurate and well written material.

But Wikipedia is also an open community of millions of people, and several researchers are already taking advantage of it to study social interactions [2], [12], [13], [14]. An interesting level of social interaction that remains unexplored is that between the massive, egalitarian, disorganised and powerless group of editors and the small and totalitarian but much more powerful and better-organised group of administrators. The fact that such interactions are different from regular relationships between equal editors was hinted in [12]; however, as of yet, no study has focused on it.

6. CONCLUSION

In this paper, we presented a survey of the impact Wikipedia has had so far on scientific research. We began by presenting a representative sample of projects that use Wikipedia in several branches of NLP, knowledge acquisition, social studies, and even some studies about Wikipedia itself. We then studied the main arguments for and against using Wikipedia and finished by anticipating future research directions as scientists learn to exploit more of the semantic information encoded in its structure. Indeed, although Wikipedia has a number of non-negligible flaws, we believe that its positive aspects are much more considerable. As a consequence, we expect its use in scientific research to become more common and more accepted over time, and that at some point in the future it will become a standard resource.

REFERENCES

[1] Schönhofen, P. (2006) "Identifying document topics using the Wikipedia category network", *Proc. of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pp. 456-462.

[2] Brandes, U. & Lerner, J. (2007) "Visual analysis of controversy in user-generated encyclopedias", *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*, pp. 179-186.

[3] Nastase, V. & Strube, M. (2008) "Decoding Wikipedia Categories for Knowledge Acquisition", *Processing of the Twenty-Third AAAI Conference on Artificial Intelligence*, pp. 1219-1224.

[4] Wee, L. C. & Hassan, S. (2008) "Exploiting Wikipedia for directional inferential text similarity", *Proceedings of the Fifth International Conference on Information Technology: New Generations*, pp. 686-691.

[5] Ahn, D., Jijkoun, V., Mishne, G., Müller, K., de Rijke, M., & Schlobach, S. (2005) "Using Wikipedia at the TREC QA track", *Proceedings TREC 2004*.

[6] Bunescu, R. & Pasca, M. (2006) "Using encyclopedic knowledge for named entity disambiguation", *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*.

[7] Voss, J. (2005) "Measuring Wikipedia", in *Proceedings of the International Conference of the International Society for Scientometrics and Informetrics*.

[8] Sorg, P., & Cimiano, P. (2008) "Cross-Lingual Information Retrieval with Explicit Semantic Analysis", *Working Notes for the CLEF 2008 Workshop*.

[9] Milne, D., Medelyan, O. & Witten, I. H. (2006) "Mining domain-specific thesauri from Wikipedia: a case study", *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pp. 442-448.

[10] Wu, F., & Weld, D.S. (2008) "Automatically Refining the Wikipedia Infobox Ontology", *Seventeenth International World Wide Web Conference*, Beijing, China.

[11] Wang, P., Hu, J., Zeng, H.-J., Chen, L., & Chen, Z. (2007) "Improving text classification by using encyclopedia knowledge", *Proceedings of the Seventh IEEE International Conference on Data Mining*, pp. 332-341.

[12] Suh, B., Chi, E. H., Pendleton, B. A., & Kittur, A., (2007) "Us vs. them: understanding social dynamics in Wikipedia with revert graph visualizations", *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*, pp. 163-170.

[13] Viégas, F. B., Wattenberg,, M. & Dave, K. (2004) "Studying cooperation and conflict between authors with history flow visualization", *Proceedings of the Conference of the ACM Special Interest Group on Computer-Human Interaction*, pp. 575-582.

[14] Viégas, F. B., Wattenberg, M., Kriss, J., & van Ham, F. (2007) "Talk before you type: coordination in Wikipedia", *Proceedings of the 40th Annual Hawaii International Conference on System Sciences*, pp. 78-78.

[15] Thomas, C. & Sheth, A. P. (2007) "Semantic convergence of Wikipedia articles", *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, pp. 600-606.

[16] Lim, E.-P., Vuong, B.-Q., Lauw, H. W., & Sun, A. (2006) "Measuring qualities of articles contributed by online communities", *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pp. 81-87.

[17] Viégas, F. B. (2007) "The Visual side of Wikipedia", *Proceedings of the 40th Annual Hawaii International Conference on System Sciences*, pp. 85-85.

[18] Buriol, L. S., Castillo, C., Donato, D., Leonardi, S., & Millozzi, S. (2006) "Temporal analysis of the Wikigraph", *Proc. of the 2006 IEEE/WIC/ACM Int'l Conf. on Web Intelligence*, pp. 45-51.

[19] Francis, W. N. & Kučera, H. (1964) *Manual of information to accompany a standard corpus of present-day edited American English, for use with digital computers*, Department of Linguistics, Brown University, Providence, Rhode Island.

[20] Read, B. (2006) "Can Wikipedia ever make the grade?", *The Chronicle of Higher Education: Information Technology*, Volume 53, Issue 10, Page A31.

[21] Rosenzweig, R. (2006) "Can History be Open Source? Wikipedia and the Future of the Past", *The Journal of American History*, Volume 93, Number 1, pp. 117-146

[22] Greenstein, S. (2007) "Wagging Wikipedia's long tail", *IEEE Micro*, Vol. 27, Issue 2, pp. 6-6.

[23] Svoboda, E. (2006) "One-click content, no guarantees", *IEEE Spectrum*, Vol. 43:5, pp. 64-65.

[24] Giles, J. (2005) "Internet Encyclopaedias go head to head", *Nature*, Volume 438, pp. 900-901.

[25] Encyclopædia Britannica Inc. (2006) "Fatally Flawed: Refuting the recent study on encyclopedic accuracy by the journal Nature" http://corporate.britannica.com/britannica_nature_response.pdf, accessed: May 2008.

[26] Wissner-Gross, A. D. (2006) "Preparation of topical reading lists from the link structure of Wikipedia", *Proceedings of the Sixth International Conference on Advanced Learning Technologies*, pp.825-829.