INFORMATION UNIQUENESS IN WIKIPEDIA ARTICLES

Nikos Kirtsis, Sofia Stamou[1], Paraskevi Tzekou and Nikos Zotos
Computer Engineering and Informatics Department, Patras University 26500 Greece
[1]Department of Archives and Library Science, Ionian University, 49100 Greece

Keywords: Wikipedia, Information Uniqueness, Content Duplication, Quality Assessment.

Abstract: Wikipedia is one of the most successful worldwide collaborative efforts to put together user generated content in a meaningfully organized and intuitive manner. Currently, Wikipedia hosts millions of articles on a variety of topics, supplied by thousands of contributors. A critical factor in Wikipedia's success is its open nature, which enables everyone edit, revise and /or question (via talk pages) the article contents. Considering the phenomenal growth of Wikipedia and the lack of a peer review process for its contents, it becomes evident that both editors and administrators have difficulty in validating its quality on a systematic and coordinated basis. This difficulty has motivated several research works on how to assess the quality of Wikipedia articles. In this paper, we propose the exploitation of a novel indicator for the Wikipedia articles' quality, namely *information uniqueness*. In this respect, we describe a method that captures the information duplication across the article contents in an attempt to infer the amount of distinct information every article communicates. Our approach relies on the intuition that an article offering unique information about its subject is of better quality compared to an article that discusses issues already addressed in several other Wikipedia articles.

1 INTRODUCTION

Wikipedia is one of the most popular social media web sites that allows users to create content in a collaborative manner. As of October 2009, the English Wikipedia hosts over 3 million[1] articles and it keeps growing as new material is being supplied by numerous editors who work on a volunteer basis. The open participation that Wikipedia offers to editors has lead to its remarkable growth but at the same time it has raised doubts about the quality of its contents, considering that anyone can freely create new or modify existing content.

In light of its success the question of how to asess the Wikipedia articles' quality is of paramount importance. To that end, several methods have been proposed for measuring Wikipedia's quality (Stvilia, et al., 2005a; Blumenstock, 2008a), most of which rely on the examination of the articles' internal characteristics, such as their contextual elements (Stvilia, et al., 2005b), their linkage in the Wikipedia graph (Kamps and Koolen, 2009), their length (Blumenstock, 2008b), their factual accuracy (Giles, 2005), the formality of their language (Emigh and Herring, 2005) and many more.

Additionally, Wikipedia has launched detailed editing and linking guidelines[2] for perspective editors and has also set precise specifications about what should and what should not be part of an article. Existing quality assessment methods combined with Wikipedia editing instructions could be integrated into a generalized validation mechanism that would assist administrators detect articles that need be repaired, enriched or modified to reach good quality levels.

One issue that has not been explicitly addressed in existing Wikipedia quality assessment techniques is the uniqueness of information that Wikipedia articles communicate for their subjects. Until now, an article's quality is determined in isolation from other article contents and as such it is error-prone to misleading quality judgments. To tackle this, we propose a novel indicator of quality for the article contents, namely *information uniqueness*. To capture information uniqueness, we present a method that quantifies the information shared across Wikipedia

[1] http://en.wikipedia.org/wiki/Special:Statistics.

[2] http://en.wikipedia.org/wiki/Wikipedia:How_to_edit_a_ page.

Kirtsis N., Stamou S., Tzekou P. and Zotos N.
INFORMATION UNIQUENESS IN WIKIPEDIA ARTICLES.
DOI: 10.5220/0002841401370143
In *Proceedings of the 6th International Conference on Web Information Systems and Technology (WEBIST 2010)*, page
ISBN: 978-989-674-025-2

entries and discriminates articles of unique information about their discussed topics from articles that duplicate information contained in other Wikipedia entries of relevant topics. The innovation of our approach is that we assess an article’s quality in context with other topically relevant articles, based on the intuition that a qualitative article should discuss new information that is not part of other articles, rather than reproduce the contents of other entries.

In the course of our study, we make the following contributions:

- We quantify information uniqueness in the contents of Wikipedia articles. In this respect, we rely on articles pertaining to related topics, as these are determined by Wikipedia editors, and we examine the degree to which the articles’ body is contained (i.e., duplicated) in the body of other topically-relevant articles. Articles’ textual containment is quantified based on the articles’ overlapping lexical and semantic elements. Our computations help us infer the articles’ contextual completeness and quality and may prove useful towards establishing validation policies for the article contents.
- We estimate information uniqueness between Wikipedia articles and the external (non-wiki) sources to which they point. In particular, we compute the amount of *new* content external resources add to the articles’ body. Our computations help us infer the appropriateness of external links to complementing their corresponding article contents and may prove useful towards establishing the Wikipedia external links' update policy.

The remainder of the paper is organized as follows. We start our discussion with an overview on related work. In Section 3, we describe how information uniqueness can serve as implicit indicator of quality for the Wikipedia article contents. In Section 4, we present an experiment we carried out in which we assessed the amount to which Wikipedia articles communicate distinct and complete information about their discussed subjects. Experimental results shed light on the article features to be considered in future Wikipedia quality assessment efforts. We conclude the paper in Section 5.

2 RELATED WORK

There has been a large body of work on how to assess the quality of Wikipedia articles. In this respect, researchers have suggested a number of article features that signify quality, such as their survival period (Cross, 2006), the number and frequency of their edits (Wilkinson and Huberman, 2007), their revision history (Adler and de Alfaro, 2007), the amount of outbound citations to 'trusted' material, e.g. scientific publications (Nielsen, 2007), the dedication of their editors (Riehle, 2005) and so forth. A complete analysis of the Wikipedia features is provided in (Voss, 2005).

Besides studying the article contextual elements that signal quality, researchers have also explored the properties and evolution of the Wikipedia link graph for assessing the impact of global and local link topology structure on information retrieval tasks (Buriol et al., 2006; Koolen and Kamps, 2009). The above studies reveal two interesting trends regarding Wikipedia’s link structure: first that Wikipedia links are good indicators of relevance for the article topics and second that there is a strong correlation between the link degree and the document length in the Wikipedia collection.

Although our work relates to the above studies that investigate the Wikipedia article contents and links for inferring their quality, it is different in two significant ways. First, unlike previous works that estimate the quality of every article individually, we study the quality of an article in context with other Wikipedia articles of relevant topics. Moreover, in contrast to existing works that confine their investigations to the articles’ contextual elements, we also examine the impact of external links to their corresponding articles’ quality. In our investigations, we estimate the information uniqueness in the articles’ referential resources, in order to assess the value external links add to the article contents. Overall, we believe that our study complements existing works on measuring the quality of Wikipedia articles.

3 INFORMATION UNIQUENESS

In this section, we present our approach for capturing the contextual and referential uniqueness across topically-related Wikipedia articles. We begin our discussion by justifying the motivation for our study and we continue with the description of the methods we propose for measuring information distinctiveness in Wikipedia articles.

3.1 Motivation

Wikipedia is the largest free-content online encyclopaedia. Although there are no specialized qualifications set forth in order for someone to become editor, there exist precise editing instructions with respect to what should or could be part of an article, how to organize the article contents and how to provide links from Wikipedia articles to external (outside Wikipedia) resources. The quest in establishing structural and contextual editing guidelines is to help Wikipedia contributors supply content that is reliable, neutral, verifiable via cited sources, complete and useful for the covered subjects. Above all, it is critical that articles communicate encyclopaedic knowledge that would distinguish Wikipedia as an encyclopaedia from other information sources.

However, Wikipedia's open nature, lacking coordinated editing, is susceptible to transforming its hosting articles from information sources to information loops. This is mainly because editors might rely on common resources for editing articles on similar topics, or they might link articles to external resources that have borrowed content from Wikipedia. In the former situation, **different articles might contain identical textual fragments**, which if reproduced across several articles might result in severe content duplication, which in turn would engage users to reading the same text for different topics; thus entering in information loops. Considering the findings of (Buriol et al., 2006) that topically-relevant Wikipedia articles are densely linked, the above scenario is likely to occur as Wikipedia content and links proliferate. In the second situation, **an article might reproduce verbatim the body of its linked external sources** and vice versa. In this case, readers visiting refereed material for acquiring supplemental information for the article topics would end up re-reading the same text (or pieces of it).

Evidently, both situations hinder users from obtaining unique information in the contents of different articles and harm the overall Wikipedia quality. Driven by the desire to capture information uniqueness across Wikipedia articles, we carried out the present study. The aim of our work is to explore the information loops in Wikipedia articles in order to deduce the amount of unique information in their contents. We believe that the findings of our study will give useful insights regarding the articles' quality and will assist Wikipedia administrators determine effective article revision and maintenance policies. In the following paragraphs, we discuss the details of our approach for quantifying information uniqueness in Wikipedia articles.

3.2 Information Uniqueness in Article Contents

To capture information uniqueness in the contents of Wikipedia articles, we rely on articles dealing with related topics and we compute the degree to which different articles duplicate the same informational extracts in their body. Our speculation is that the degree of the articles' information uniqueness is conversely analogous to the degree of their content duplication, in the sense that the more content two articles share in common the less unique information they communicate.

The method we employ for estimating information uniqueness in the Wikipedia collection operates upon topically-related articles. The reason for focusing on topically related articles is because information duplication is mainly pronounced for documents dealing with common or similar subjects (Davison, 2000). Therefore, trying to capture content duplication across the entire Wikipedia collection would significantly increase the overhead and the computational complexity of our measurements without adding much value to the delivered results.

To identify topically-related articles, we rely on the Wikipedia categories to which every article has been assigned by its editors and we deem articles to be topically-related if they share at least one common category. By deducing the articles' topical relatedness based on their assigned categories, we ensure both consistency and accuracy in their topical descriptions as the latter are collectively supplied by humans and we obviate the need to re-categorize the articles.

Based on the article categories, we organize them into topical clusters and we process the documents in every cluster in order to identify duplicate content in their body. Having organized Wikipedia articles into topical clusters, we download the contents of every cluster, we parse them to remove mark-up and apply tokenization to identify word tokens in the articles' textual body. Based on the tokenized articles in every topic, we estimate the articles' lexical and semantic elements duplication from which we subsequently derive the amount of the articles' information uniqueness.

To identify lexical content duplication across the articles' body, we estimate Containment within the articles' text. For our measurements, we firstly lexically analyze the articles' textual body into canonical sequences of tokens and represent them as contiguous sub-sequences of *w* tokens, called shingles and computed via the *w*-shingling technique (Broder et al., 1997). Then, we eliminate identical shingles from every article and we compute the containment of an article's text in the body of the remaining articles clustered in the same topic. Containment between article pairs is determined as the ratio of an article's shingles that are contained in the shingles of another article, given by:

Where $S(a_i)$ denotes the shingles of article a_i and $S(a_j)$ denotes the shingles of article a_j. Containment values range between 0 and 1; with 0 indicating that the two articles share no text in common and 1 indicating that the textual body of an article in contained (i.e., appears identical) in the body of another article. Based on the above, we compute the amount of text in every article that is contained in the body of other articles, as given by:

$$Containment\left(a_i, a_j\right) = \frac{\left|S\left(a_i\right) \cap S\left(a_j\right)\right|}{\left|S\left(a_i\right)\right|} \qquad (1)$$

$$Containment\left(a_i\right) = \frac{1}{|N|}\sum_{a_i \in N}^{|N|} Containment\left(a_i, a_j\right), \ldots, \left(a_i, a_n\right) \qquad (2)$$

Where *N* indicates the number of articles considered. Eventually, we derive the lexical uniqueness of an article based on the amount of the article shingles that are not contained in other articles, i.e., they are unique among the shingles generated for the articles in a topic. Formally, the lexical uniqueness *LU* of an article is quantified as:

$$LU\left(a_i\right) = 1 - Containment\left(a_i\right) \qquad (3)$$

Based on the above, we compute uniqueness in the articles' lexical contents, so that the less text of an article is contained in the body of other articles the increased its lexical uniqueness. However, lexical uniqueness alone does not suffice for capturing information uniqueness given that articles might use distinct surface terms to verbalize the same information. To tackle such cases, we need to capture the semantic uniqueness in the article contents. In other words, we need to estimate the degree to which different articles contain semantically equivalent elements in their body and then infer their semantic uniqueness by relying on the amount of their unshared semantic information.

To estimate the semantic uniqueness of an article's content, we firstly need to annotate every word token of an article with a suitable sense. For assigning sense labels to the articles' words, we utilize WordNet (Fellbaum, 1998) as our sense repository and operate as follows. We map all words of an article to their corresponding WordNet synsets. Word tokens matching a single synset are annotated with that synset's gloss (i.e., sense), whereas words matching several synsets are labelled with the sense that exhibits the maximum average similarity to the senses identified for the remaining article words. In our work, we estimate semantic similarity based on the Wu and Palmer (1994) metric. Having annotated every word of an article with an appropriate sense, we quantify the semantic equivalence between pairs of articles as the fraction of senses they share in common, i.e.:

$$Equivalence\left(a_i, a_j\right) = \frac{\left|Senses\left(a_i\right) \cap Senses\left(a_j\right)\right|}{\left|Senses\left(a_i\right) \cup Senses\left(a_j\right)\right|} \qquad (4)$$

Where *Senses* (a_i) denotes the senses assigned to the word tokens of article a_i and *Senses* (a_j) denotes the senses assigned to the word tokens of article a_j. Equivalence values range between 0 and 1; with 0 indicating that the semantic content of the two articles is different and 1 indicating that the content of the two articles is semantically equivalent. Based on the above, we compute for every article the amount of semantic content its shares with other articles as:

$$Equivalence\left(a_i\right) = \frac{1}{|N|}\sum_{a_i \in N}^{|N|} Equivalence\left(a_i, a_j\right), \ldots, \left(a_i, a_n\right) \qquad (5)$$

Then, we estimate for every article the semantic uniqueness of its contents, based on the amount of the article semantics that are not shared in other articles. Formally, the semantic uniqueness *SU* of an article is quantified as:

$$SU\left(a_i\right) = 1 - Equivalence\left(a_i\right) \qquad (6)$$

Finally, we estimate the article's overall information uniqueness (*IU*) with respect to other topically-related articles based on the combination of the article's lexical and semantic uniqueness, given by:

$$IU\left(a_i\right) = \left(a \bullet LU\left(a_i\right)\right) + \left(\left(1 - a\right) SU\left(a_i\right)\right) \qquad (7)$$

Where *a* is a weighting factor that determines the impact of the article's lexical and semantic elements on its overall information uniqueness. Setting the value of *a*= 0.5 would weight the article's lexical and semantic properties equally, whereas lowering *a* i.e., *a* < 0.5 would promote semantic uniqueness as an indicator of the article's content distinctiveness and increasing *a* i.e., *a* > 0.5 would promote lexical uniqueness as an indicator of the article's content distinctiveness. In any case, the weighting factor ensures that information uniqueness values are normalized and range between 0 and 1; with 0 indicating total information duplication (lack of unique content) and 1 indicating total information uniqueness (lack of duplicate content).

Based on the above, we measure how distinct is the information that an article communicates for its underlying topic compared to the information contained in the body of other articles dealing with the same topic. Recall that, decreased levels of information uniqueness increase the probability of encountering information loops in the Wikipedia collection. Therefore, the more unique information Wikipedia articles contain the better quality Wikipedia bears.

3.3 Information Uniqueness in Article Referential Sources

Besides capturing information uniqueness across the Wikipedia articles, we also estimate how unique is the information contained in every article compared to the information contained in its referential external resources. Considering that Wikipedia articles provide links (refer the reader) to non-wiki sources, the information of which should be relevant and complementary to the article contents, it is useful to assess whether the information contained in the article references is part of the article contents or not. In the former case, reading the contents of the article references would enter the user to information loops and would diminish the references' added value for the article contents. In the latter case, reading the contents of the article references would point the user to *new* (not part of the article) information and would empower the references' positive contribution towards complementing the article contents.

To estimate how distinct is the information of an article with respect to the contents of its referenced sources, we start by downloading the textual body of the article's external links, which we then parse and tokenize. Afterwards, we merge together the tokenized contents of all the pages pointed by an article to formulate a super-document that we lexically analyse and represent into shingles. Having represented both the article contents and the contents of all the documents pointed by the article as shingles, we compute their lexical overlap by applying the *Containment* measure (eqn.1). Based on the amount of overlapping lexical content between an article and its referential sources, we deduce the article's *Lexical Uniqueness* (eqn.3). In addition, by semantically annotating the words contained in an article's referential sources we can estimate the fraction of senses shared between the article and its referenced sources and deduce their semantic *Equivalence* (eqn.5) upon which we rely for quantifying the article's *Semantic Uniqueness* (eqn.6) with respect to its external resources. Finally, we combine the two measurements and compute the article's *Information Uniqueness* (eqn.7) compared to the contents of its referential sources.

4 EXPERIMENTS

To capture information uniqueness in the Wikipedia collection, we relied on the October 2009 dump of the English Wikipedia, a 24.5 GB xml corpus, which we processed in order to extract the articles it contained and for every article extract its associated Wikipedia categories and the URLs of its external resources. Having processed the Wikipedia corpus, we sampled 100,000 random articles associated with common Wikipedia categories and used them as the dataset against which we would apply our information uniqueness measurements.

Based on the identified article categories, we organized the sampled articles into clusters, with every cluster grouping together articles of a common category. Then, we processed the contents of every article in each cluster to represent the articles' textual elements as shingles, i.e., contiguous sequences of *w* tokens (with *w*= 10). Based on the comparative lexical and semantic analysis of the article shingles, we quantified the information uniqueness of every article with respect to other articles organized in the same cluster. Results are presented in Section 4.1.

For our sampled articles we also downloaded the contents of their external links, which we processed as previously described. Then, we comparative analyzed the contents of every article and its corresponding external resources in order to quantify the amount of distinct information the article contains compared to the information contained in its referential sources. Obtained results are discussed in Section 4.2.

4.1 Unique Information in Wikipedia

Figure 1 plots the information uniqueness in the examined Wikipedia articles. The *x*-axis illustrates the amount of articles and the *y*-axis shows the amount of unique information in the articles' body, estimated by setting the value of *a* = 0.5.

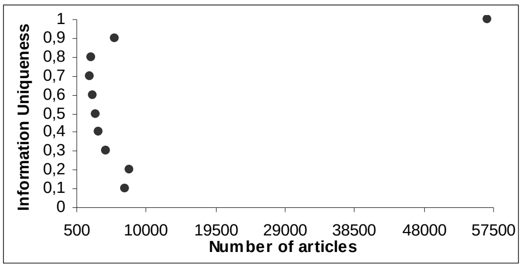

Figure 1: Unique information across sampled articles.

As the figure shows, most of the examined articles contain unique information, i.e., their textual and semantic elements are not duplicated across other articles. According to results, 72.8% on average the information in the body of all examined articles is unique and 27.2% of the information is duplicated in the body of other articles.

Although not schematically illustrated due to space constrains, a closer look at obtained results reveals that on average 88.55% of the articles' lexical elements are unique, i.e., they are not contained in the body of other articles and 74.51% of the articles' semantic content is unique. Results suggest that the information contained in most of the examined articles is unique and as such as may deduce the good overall quality of the Wikipedia corpus.

4.2 New Information brought by Wikipedia

Figure 2 plots the information uniqueness between the examined Wikipedia articles and the body of their respective linked-to external resources. The *x*-axis shows the amount of articles and the *y*-axis shows the amount of unique information the articles contain compared to the contents of their external resources.

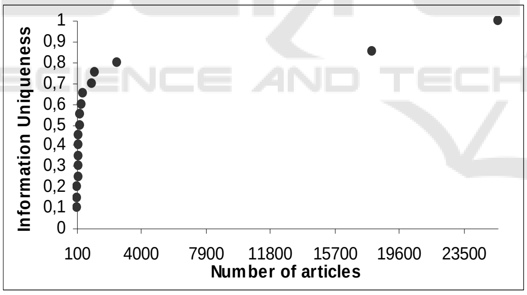

Figure 2: Unique information between the sampled articles and the contents of their external resources.

As the figure shows the information contained in most Wikipedia articles differs from the information delivered in their respective external resources. Results show that on average 89.78% of the information in all examined articles is unique compared to the information contained in the body of their linked non-wiki sources. Results suggest that the information offered to Wikipedia readers via external links adds value to the article contents as it mainly concerns *new* information that is not part of the article.

Overall, we may conclude that the vast majority of Wikipedia articles comply with the requirement that they should provide useful content that can be verified via cited external resources.

5 CONCLUDING REMARKS

In this paper, we have proposed the exploitation of information uniqueness as a novel indicator of quality for the Wikipedia articles. The application of our approach to a subset of the Wikipedia corpus revealed that most articles contain unique information about their subjects, thus signifying the overall good quality of Wikipedia. Our measurements, especially those indicating articles of high information duplication, could help Wikipedia administrators set better linking instructions between distinct articles so that editors provide links between articles instead of repeating their contents. Moreover, our findings could assist Wikipedia administrators implement contextual filtering mechanisms that would enable editors determine whether external links contain supplemental or superfluous information for the article contents. In the future, we plan to extend our preliminary study and capture information uniqueness across the entire Wikipedia collection in order to quantify its overall quality.

REFERENCES

Adler, N.T., de Alfaro, L., 2007. A content-driven reputation system for the Wikipedia. In *Proceedings of the 16th International World Wide Web Conference.*

Blumenstock, J.E, 2008a. Automatically assessing the quality of Wikipedia articles. *UBCiSchool Report*. 2008-021

Blumenstock, J.E, 2008b. Size matters: word count as a measure of quality in Wikipedia. In *Proceedings of the 17th Intl. WWW Conference*.

Buriol, J., Castillo, C., Donato, D., Leonardi, S., Millozzi, S., 2006. Temporal evolution of the wiki graph. In *Proceedings of the Web Intelligence Conference*.

Broder, A.Z., Glassman, S.C., Manasse, M.S., Zweig, G. 1997. Syntactic clustering of the web. In *Proceedings of the 6th Intl. WWW Conference*, pp. 391-404

Cross, T., 2006. Puppy smoothies: improving the reliability of open, collaborative wikis. *First Monday*, 11 (9).

Davison, D. 2000. Topical locality on the web. In *Proceedings of the 23rd Intl. SIGIR Conference*.

Emigh, W., Herring, S., 2005. Collaborative authoring on the web: a genre analysis of online enclyclopedias. In *Proceedings of the HICSS Conference*.

Fellbaum, Ch. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Giles, J., 2005. Internet encyclopaedias go head to head. In *Nature*, 438:900-901.

Kamps, J., Koolen, M., 2009. Is Wikipedia ;link structure different? In *Proceedings of the 2nd Intl. WSDM Conference*.

Koolen, M., Kamps, J., 2009. What's in a link? From document importance to topical relevance. In *Proceedings of the 2nd International Conference on Theory of Information Retrieval*, pp. 313-321.

Nielsen, F.A., 2007. Scientific Citations in Wikipedia. In *Computing Research Repository*.

Rieche, D., 2005. How and why Wikipedia works? an interview. In *Proceedings of the ACM Wikisym*.

Stvilia, B., Twidale, M.B., Smith, L.C., Gasser, L. 2005a. Assessing information quality of a community-based encyclopaedia. In *Proceedings of the International Conference on Information Quality*.

Stvilia, B., Twidale, M.B., Gasser, L., Smith, L.C.. 2005b. Information quality discussions in Wikipedia. In *Proceedings of the International Conference on Knowledge Management*.

Voss, J., 2005. Measuring Wikipedia. In *Proceedings of the 10th International Conference of the International Society for Scientometrics and Infometrics*.

Wilkinson, D.M., Huberman, B.A., 2007. Assessing the value of cooperation in Wikipedia. *First Monday*, 12(4).

Wu, Z., Palmer, M. 1994. Verb semantics and lexical selection. In *Proceedings of the 32nd ACL Conference*, pp. 133-138.