# A method for automated document classification using Wikipedia-derived weighted keywords

**Article** · March 2015

DOI: 10.1109/ICODSE.2014.7062484

CITATIONS
3

READS
266

**2 authors**, including:

Robert P. Biuk-Aghai
Software Company
**93** PUBLICATIONS **757** CITATIONS

SEE PROFILE

http://dx.doi.org/10.1109/ICODSE.2014.7062484

# A Method for Automated Document Classification Using Wikipedia-Derived Weighted Keywords

Robert P. Biuk-Aghai
Department of Computer and Information Science
Faculty of Science and Technology
University of Macau
Email: robertb@umac.mo

Ka Kit Ng
Department of Computer and Information Science
Faculty of Science and Technology
University of Macau
Email: ngkakit@gmail.com

***Abstract*—The pace of knowledge creation such as in academic research has accelerated rapidly in recent years, resulting in ever more new research publications. This has made it difficult to keep abreast of new developments, or to know which new publications are relevant to a given research area. We have developed a method for analysing and automatically classifying publications. Our method makes use of the Wikipedia category hierarchy, and the content of Wikipedia articles associated to Wikipedia categories. Initially we perform pre-processing and simplification of the Wikipedia category hierarchy, resulting in a rooted directed graph. Wikipedia articles are then analysed, and a set of keywords per Wikipedia category are extracted using a modified tf-idf (term frequency-inverse document frequency) model proposed in this paper. To classify a given input document, tf-idf weights are used to extract relevant keywords from the document, which are then matched to the keywords previously extracted from Wikipedia. The closest matching top-level categories are identified from all categories containing the document's keywords. A cosine similarity metric is then applied to select the closest matching sub-category, recursing down the category hierarchy until the best matching categories are identified. The final result produced shows a set of categories matching the input document, together with a matching percentage. This result can be used to identify new documents that are relevant to a specific research area, or to classify a whole set of documents into different topic areas, with sub-topics, main keywords, and associated weights. We present an experimental study using data from English Wikipedia.**

## I. Introduction

Academic research has witnessed an explosion in the 20th century that is unparalleled in human history and that has continued unabated into the current century. Never before has so much knowledge been created, published and disseminated as in today's age. Indeed, worldwide scientific research output steadily increases. Larsen and von Ins studied the growth of science from 1907 until 2007 as measured by the number of SCI/SSCI-indexed publications and reported that well over one million peer-reviewed journal articles are published every year, with an annual growth rate of 2.3% [1].

Similar results are reported by the National Science Board of the USA who documented an annual worldwide growth rate of science and engineering research output between 2001 and 2011 of 2.8% [2]. This was up from the 2.6% growth rate reported a decade earlier based on data from 1988 to 2001 [3].

Clearly, large amounts of academic research are being published every year, making it impossible to stay abreast of all new developments in anything but the very narrowest of research areas. Our work presented in this paper was motivated by the desire to get an overview of the existing publications related to Wikipedia research. As we soon discovered, even in such a narrow research area there are too many publications for our author team of two to review. Table I shows the number of Wikipedia-related publications indexed in different sources, by year since 2001, the year of Wikipedia's founding. Although there is some overlap between these different sources, and potentially some duplication within Google Scholar, in most recent years there are over 100 publications, and for all years and sources combined probably well over 1000 publications. A researcher who is new to this research area will have difficulty even selecting the most relevant papers, let alone manually reviewing them all. Thus we devised a method to automatically process a collection of research papers, to classify each paper as belonging to certain topic categories, and to extract its most significant keywords. When applying this method to a whole collection of papers for a research area one can obtain a high-level overview of the different directions and sub-areas, which can greatly help a researcher orient oneself in that area.

TABLE I. Number of publications with "Wikipedia" in the document title, as of 1 Oct 2014 (search scope: ACM DL: within The ACM Guide to Computing Literature; IEEExplore: entire collection; Web of Science: Core Collection; Google Scholar: all, excluding patents and citations)

| Year | ACM DL | IEEE xplore | Web of Science | Google Scholar |
|---|---|---|---|---|
| 2001 | 0 | 0 | 0 | 7 |
| 2002 | 0 | 0 | 1 | 2 |
| 2003 | 0 | 0 | 1 | 4 |
| 2004 | 0 | 0 | 0 | 14 |
| 2005 | 4 | 0 | 10 | 38 |
| 2006 | 16 | 5 | 21 | 142 |
| 2007 | 27 | 12 | 50 | 228 |
| 2008 | 47 | 32 | 95 | 377 |
| 2009 | 128 | 25 | 128 | 432 |
| 2010 | 137 | 45 | 102 | 548 |
| 2011 | 149 | 39 | 107 | 548 |
| 2012 | 137 | 31 | 93 | 563 |
| 2013 | 105 | 26 | 115 | 508 |
| 2014 | 48 | 13 | 43 | 277 |
| **Total** | 798 | 228 | 766 | 3661 |

In the following section we briefly review some related work, and then introduce our method in Section III. In Section IV we present an experimental evaluation of our method, and show an application of this method in Section V, before making conclusions in Section VI.

## II. Related Work

The use of Wikipedia data for constructing ontologies, extracting keywords, and mapping topics has increased parallel to the growth in scope and increase in quality of Wikipedia content. Whereas in the first few years of its existence Wikipedia's editing process, which allows anyone to edit any article, was criticized by publishers of traditional encyclopedias for not assuring the quality and reliability of content, there has been growing acknowledgement of the high quality of many Wikipedia articles [4], in some cases rivalling that found in other encyclopedias. In this connection, much work has focused on both measuring [5], [6], [7], [8], [9] and improving [10], [11] Wikipedia article quality.

Related to document classification there are several cases of related work that use methods belonging to text mining, keyword extraction and machine learning. One example is the use of the Conditional Random Fields (CRF) model which has been applied to the problem of keyword extraction and found to perform very well compared with other methods [12], [13]. The work of Coursey et al. [14], [15] has focused on topic identification using a biased graph centrality algorithm applied to Wikipedia data. It identifies topics relevant to a given document, based on a graph of Wikipedia articles and categories.

Wikipedia data has been used to construct an ontology, which has been used to tag documents with related terms [16]. This approach is closest to the one we propose in this paper. However, our approach differs by using weighted keywords to determine related topics.

## III. Method

We propose a method for automatically classifying documents in one or more categories. Inputs are the Wikipedia category system and a document (such as a research paper) to be classified. Outputs are the categories that this paper belongs to, and a set of keywords with matching score.

In outline, our method operates by extracting a set of descriptive keywords from the paper to be classified, and attempting to match these with descriptive keywords extracted from each category in Wikipedia. The closest match among these sets of keywords indicates the category that best describes the paper.

Operationally our method consists of two main phases: (1) a pre-processing phase, and (2) a classification phase, each of which consists of several steps. In the pre-processing phase the Wikipedia category graph is analysed and a set of keywords is generated for each category. In the classification phase the paper to be classified is analysed and a set of keywords is generated, and subsequently the best matches among this set of keywords and those from Wikipedia is obtained. The entire process is illustrated in Figure 1.

### A. Phase 1: Pre-processing

The pre-processing phase takes four inputs: (1) the set of all Wikipedia articles, (2) the set of all Wikipedia categories, (3) the set of all category-to-category (parent-child) relationships, and (4) the set of all article-to-category relationships.

#### *Step 1.1: Pre-process Wikipedia category graph*

The Wikipedia category graph contains a few anomalies such as cycles. To facilitate later processing of the graph, we pre-process it to eliminate these anomalies. The result is a rooted directed acyclic graph such as the extract from English Wikipedia displayed in Figure 2 (data of 1 Oct 2014). This figure shows the root of the category graph (category "Contents") under which the entry point into the categories of all Wikipedia articles is the category "Articles". On the next level down English Wikipedia uses two systems of categorisation, of which we use the one rooted at category "Main topic classifications" as it contains a more detailed sub-division into further categories (a total of 22 categories on the next level down, out of which for space reasons we only show the three categories "History", "Science" and "Technology"). Nodes in the category graph may have multiple parents, such as the category "History of science" which has two parents "History by topic" and "Science"; and the category "History of technology" which has three parents "History by topic", "History of science" and "Technology" (plus several other parent categories not shown). Many thousands of other categories have been omitted in this figure, and the grey arrows and dots indicate that the graph extends further in depth and breadth than what is pictured.

#### *Step 1.2: Extract category keywords from Wikipedia*

Each category in Wikipedia can be represented by a number of keywords that best describes it. However, the Wikipedia database does not provide these keywords, so we have to extract them. For each Wikipedia category we retrieve all articles assigned to that category. For each article we then perform stemming and stop word removal on the article text. Next we extract keywords for each article by computing tf-idf (term frequency-inverse document frequency) weights [17] according to (1).

$$w\left(t,c\right) = tf\left(t,c\right) \times \log\left(\frac{N}{n}\right) \tag{1}$$

where $w(t,c)$ is the weight for term $t$ in the collection of articles belonging to category $c$; $tf(t,c)$ is the frequency of term $t$ in the collection of articles belonging to category $c$; $N$ is the total number of articles in category $c$; and $n$ is the number of articles in $c$ in which term $t$ occurs.

The result of the tf-idf weight computation is a set $T_c$ of terms with associated weights $w(t,c)$ for each category $c$. From each set $T_c$ we eliminate all but the top-scoring $x$ terms (i.e. terms with the highest weights). Parameter $x$ is empirically determined. The final outcome of this step is a set of category keywords $K_c$.

#### *Step 1.3: Weight keywords on parent-child category links*

The Wikipedia category system is a graph of related categories, and as observed above there are many instances of categories belonging to multiple parent categories. Moreover, many category keywords are shared among multiple categories. We wish to determine not only which specific category a document matches, but also the path of parent-child category relationships that best matches the document. For a category with multiple parents, any keywords found in

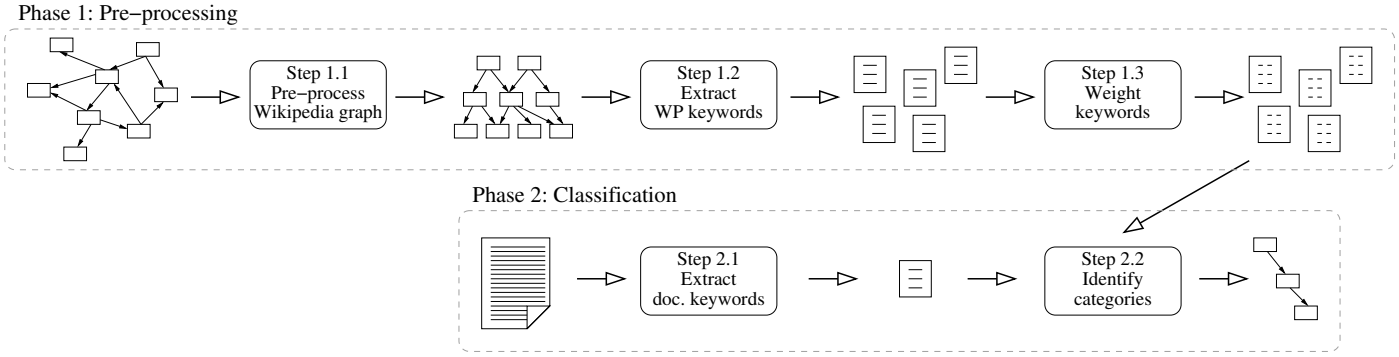


Fig. 1. Process of pre-processing and classification

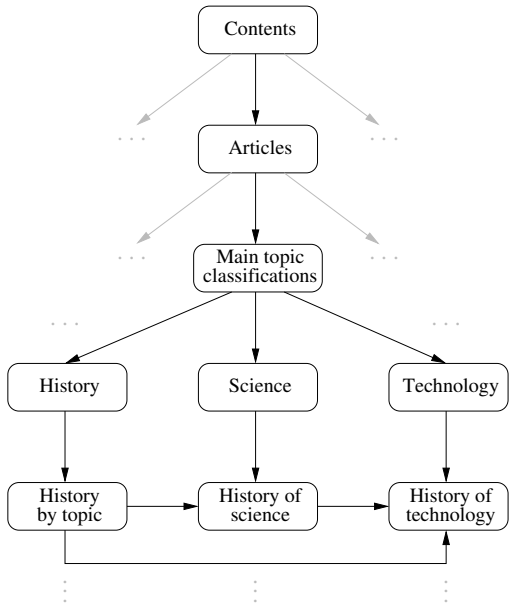


Fig. 2. Extract from the Wikipedia category graph

both parent and child categories may belong more strongly to one parent category than the other. For example, in Figure 2 we saw category "History of technology" belonging to parent categories "History of science" and "Technology". However, a given set of keywords extracted from the category "History of technology" may more closely match "Technology" than "History of science", and in describing a given document it would therefore be more suitable to identify it as belonging to the path "Technology" → "History of technology".

Therefore we weigh keywords using each parent category's keywords. For each pair of categories we obtain the weights for their shared set of keywords using a modified form of the tf-idf computation, as given in (2).

$$w\left(k,c\right)=\frac{tf\left(k,c\right)\times\log\left(\frac{N}{n}\right)}{\max\left(tf\left(k_i,c\right)\times\log\left(\frac{N}{n}\right)\right)} \tag{2}$$

This calculation is performed for all shared keywords $k \in K_c, k \in K_{pc}$ ($K_c$ is the set of keywords of a given category $c$, $K_{pc}$ is the set of keywords of its parent category $pc$). The result is normalized (using $max(\ldots)$ for all $i$ shared keywords) to enable comparison among different category pairs.

TABLE II. TF-IDF SCORES FOR KEYWORDS OF A CATEGORY (SPORTS BUSINESS) WITH TWO DIFFERENT PARENT CATEGORIES (BUSINESS AND SPORTS)

| **Business → Sports Business** | | **Sports → Sports Business** | |
|---|---|---|---|
| **Keyword** | **tf-idf** | **Keyword** | **tf-idf** |
| camp | 40.88 | camp | 274.50 |
| yale | 30.28 | yale | 170.08 |
| port | 54.36 | port | 79.97 |
| donut | 62.21 | donut | 62.21 |
| private | 30.74 | private | 244.83 |
| merger | 1104.29 | merger | 25.76 |
| . . . | . . . | . . . | . . . |
| Average | 220.46 | | 142.89 |

An example is illustrated in Table II for the category "Sports Business" which has two parent categories "Business" and "Sports". Tf-idf weights for keywords of category "Sports Business" that also occur in its parent category are calculated and averaged. The higher average indicates a closer relationship, in this case between "Sports Business" and "Business". This averaged weight is assigned to the edge connecting these two categories in the category graph.

### *B. Phase 2: Classification*

The classification phase takes three inputs: a given document to be classified, the Wikipedia category graph with assigned weights, and the set of keywords for each category.

*Step 2.1: Extract document keywords*

Similar to step 1.2 above, we obtain a number of keywords that best describes a given document. We perform stemming and stop word removal on the document text. Next we extract keywords for the document.

*Step 2.2: Identify related categories*

Given the set of document keywords, and the sets of category keywords from phase 1, we can now classify the document into the most related categories. This is done following the Wikipedia category hierarchy, starting with the 22 top-level categories directly under the category "Main topic classifications". For each of these categories, the similarity between the set of document keywords and the set of category keywords is computed using a cosine similarity measure, as in (3).

$$sim\,(d,c) = \frac{\Sigma_{i=1}^{t} k_{di} \times k_{ci}}{\sqrt{\Sigma_{i=1}^{t} (k_{di})^2 \times \Sigma_{i=1}^{t} (k_{ci})^2}} \quad (3)$$

Here $d$ is the given document, $c$ a Wikipedia category, and $k$ is a keyword belonging to document $d$ or category $c$. The resulting similarity value is in the range $[0, 1]$, with a higher value indicating a greater similarity. Cosine similarity is a standard measure used in information retrieval and data mining, considering each document as a vector of terms. Words occurring in a document are its terms in the vector, and all documents' vectors are part of a high-dimensional vector space. The cosine similarity measures the cosine of the angle between any pair of vectors, which is 1 in the case of two vectors with the same orientation, meaning maximum similarity; and 0 in the case of two vectors perpendicular to each other, meaning maximum dissimilarity.

Once similarity measures have been calculated for all categories at a given level, the given source document $d$ can be classified into the best matching category, i.e. the category with the highest similarity score. This process is then repeated on the next level down using the best matching category's child categories, and repeatedly on further levels down, until the highest similarity score from the child category level is less than that of its parent. This is based on the observation that the highest similarity score of a set of categories on a given level typically increases with the level until it reaches a maximum, after which it decreases. The result is a path of Wikipedia categories matching the given document. The process of classification is shown in Algorithm 1.

**Algorithm 1** Classification algorithm

```
1:  function CLASSIFY(d)
2:      path ← ∅
3:      maxSim ← 0
4:      C ← all top-level categories
5:      repeat
6:          parentMax ← maxSim
7:          maxSim ← 0
8:          for all c ∈ C do
9:              s ← sim(d, c)
10:             if s > maxSim then
11:                 maxSim ← s
12:                 maxCat ← c
13:             end if
14:         end for
15:         if maxSim ≥ parentMax then
16:             path ← append(path, maxCat)
17:             C ← child categories of maxCat
18:         end if
19:     until maxSim < parentMax
20:     return path
21: end function
```

## IV. Evaluation

To evaluate our method we applied it to a collection of research papers and manually verified the classification results. The papers were obtained by searching Google Scholar with the search term "Wikipedia" and the name of one of the 24 top-level Wikipedia categories (today there are only 22 top-level categories, but at the time of our evaluation there existed 24 such categories in Wikipedia). For each of these 24 result sets we retrieved the first 60 results for which the corresponding full-text document was available. This yielded a total of 1440 research papers. We then applied our method to those documents, obtaining the highest matching path of categories that each document is classified as belonging to. Finally we manually inspected the result to verify the automatic classification, rating it as a strong, medium, or poor match.

An extract of the results for top-level category Technology (i.e. with search terms "Wikipedia" and "Technology") is shown in Table III. Except for the first research paper, each paper is categorised into a category path three levels deep, which results in a fine-grained and descriptive classification. Some papers are in similar but not identical research areas, and our method categorises them accordingly into a different lower-level category. For example, papers 5 and 7 are both in the research area of natural language processing (category path Language → Natural language and computing), but at the third category level are assigned to different sub-categories. Similarly, papers 3 and 6 also share the same first two levels of their category path (Language → Linguistics), but are more finely distinguished at the third category level. Verifying the classification result, we found it to be a strong match for nine out of the shown ten research papers. Only one paper was evaluated as a medium match, having been categorised under Language → Linguistics → Semantics, which would have been better categorised under Science → Methodology → Scientific method. Overall, for our experiment we found a strong match in 89.1% of cases, which is a very high accuracy and demonstrates the effectiveness of our method.

## V. Application

We have implemented our method in a software prototype. The software is web-based and lets the user upload a document to be analysed, then determines the best matching categories. The result is shown in summary as matching percentages per category, supplemented by a chart displaying these percentages graphically. Figure 3 shows an example of the category matching summary for the first, second and third category levels for the paper *Emerging ICTs and their potential in revitalizing small scale agriculture in Africa*. On the first category level the highest matching percentage is in the Agriculture category; this category is thus selected for analysis on the next level where category Rural society has the highest score; on the third level this category is further analysed and category Rural culture is identified with the highest matching score.

Besides the matching summary, our application also displays the top matching keywords, their matching score, and extracts from the source document showing the keyword in context. Figure 4 shows an example of this corresponding to the two top-scoring keywords of the first level category classification summary shown above.

## VI. Conclusion

We have presented a new method for automatically classifying documents into categories taken from the Wikipedia category system. This offers a rich classification system for

TABLE III. EXPERIMENTAL RESULT OF CLASSIFYING RESEARCH PAPERS INTO WIKIPEDIA CATEGORIES

| # | Document Title | Category Path | Match |
|---|---|---|---|
| 1 | Young Adults' Credibility Assessment of Wikipedia | Education → Philosophy of education | Strong |
| 2 | Introducing New Features to Wikipedia: Case Studies for Web Science | Science → Methodology → Open methodologies | Strong |
| 3 | Exploiting Wikipedia as External Knowledge for Named Entity Recognition | Language → Linguistics → Applied linguistics | Strong |
| 4 | Learning to Trust the Crowd: Some Lessons from Wikipedia | Technology → Projects → Collaborative projects | Strong |
| 5 | WikiBABEL: A System for Multilingual Wikipedia Content | Language → Natural language and computing → Computer-assisted translation | Strong |
| 6 | Knowledge Derived from Wikipedia for Computing Semantic Relatedness | Language → Linguistics → Semantics | Medium |
| 7 | Large-Scale Named Entity Disambiguation based on Wikipedia Data | Language → Natural language and computing → Computational linguistics | Strong |
| 8 | Enacting Social Argumentative Machines in Semantic Wikipedia | Technology → Science and technology studies → Social epistemology | Strong |
| 9 | Overcoming the Brittleness Bottleneck using Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge | Science → Methodology → Learning methods | Strong |
| 10 | Normative Behaviour in Wikipedia | Society → Communication → Communication theory | Strong |

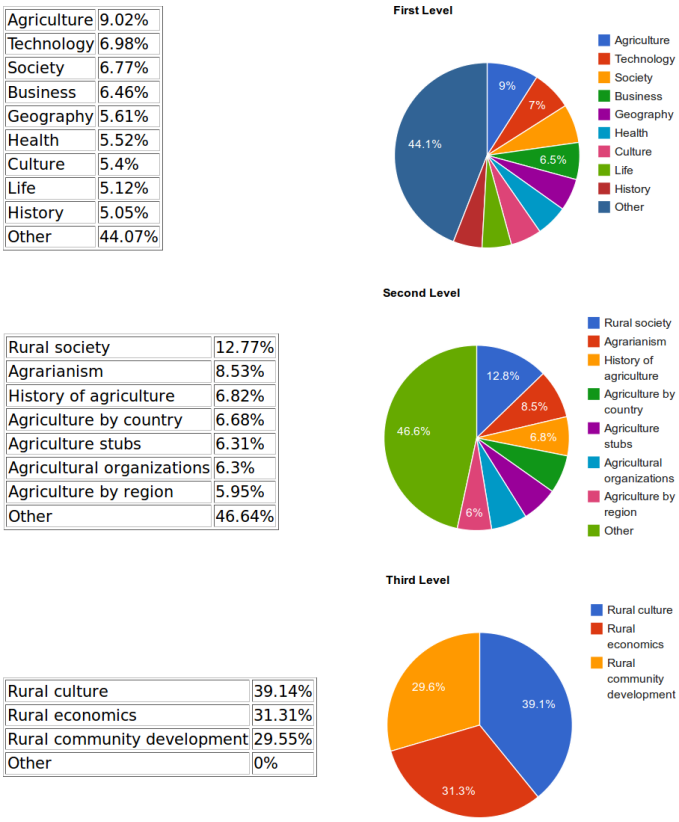


Fig. 3. Category classification summary

First level top keywords:

**1. school (34.28%) :**

(1) - College and high **school** students currently represent the largest population of Internet users (Eastin, 2001) making them an important subset to study .

(2) - The Pew Internet Project noted that 87% of all 12 to 17 year olds use the Internet, and 78% of them use it at **school** (Rainie & Hitlin, 2005) .

(3) - It has been stated by some that most college freshmen will appear on campus with newer technology than many of the **schools** themselves have (∗Freshmen Arrive§, 2006) .

(4) - A recent study of teenagers noted Proceedings of the 2007 AAAE Research Conference, Volume 34 283 that 71% use the Internet as a primary source for **school** projects, effectively replacing .

**2. student (20.15%) :**

(1) - Many college **students** have described the Internet as a functional tool that helps them to communicate with professors, do research, and access library materials .

(2) - As more and more **students** and educators are envisioning the Internet as a source for information to be used in the classroom, it is important that we monitor **students** ∗ attitudes and usage .

(3) - As educators we need to ensure **students** know how to use technology effectively by recognizing credible sources and utilizing the correct technology for each situation .

(4) - This study utilizes a descriptive survey to understand the current usage and attitudes toward the Internet by **students** enrolled in college of agriculture courses at a large Southeastern Land-Grant University .

Fig. 4. First level category top keywords, matching score, and document extracts

fine-grained identification of the actual content of a document. In contrast, commercial indexing services categorise publications in a relatively much smaller number of categories (such as the Web of Science which uses 251 categories across all topic areas of sciences, engineering, humanities, arts, medicine, etc.). Theirs is a much more rough categorisation than ours which is based on thousands of Wikipedia categories.

## REFERENCES


[1] P. Larsen and M. von Ins, "The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index," *Scientometrics*, vol. 84, no. 3, pp. 575–603, 2010. [Online]. Available: http://dx.doi.org/10.1007/s11192-010-0202-z

[2] National Science Board, "Science and engineering indicators 2014," National Science Foundation, Arlington, VA, USA, Tech. Rep. NSB 14-01, 2014. [Online]. Available: http://www.nsf.gov/statistics/seind14/

[3] ——, "Science and engineering indicators 2004," National Science Foundation, Arlington, VA, USA, Tech. Rep. NSB 04-01, 2004. [Online]. Available: http://www.nsf.gov/statistics/seind04/

[4] J. Giles, "Internet encyclopaedias go head to head," *Nature*, vol. 438, pp. 900–901, 2005.

[5] M. Hu, E.-P. Lim, A. Sun, H. W. Lauw, and B.-Q. Vuong, "Measuring article quality in Wikipedia: Models and evaluation," in *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, ser. CIKM '07. New York, NY, USA: ACM, 2007, pp. 243–252. [Online]. Available: http://doi.acm.org/10.1145/1321440.1321476

[6] T. Wöhner and R. Peters, "Assessing the quality of Wikipedia articles with lifecycle based metrics," in *Proceedings of the 5th International Symposium on Wikis and Open Collaboration*, ser. WikiSym '09. New York, NY, USA: ACM, 2009, pp. 16:1–16:10. [Online]. Available: http://doi.acm.org/10.1145/1641309.1641333

[7] G. De la Calzada and A. Dekhtyar, "On measuring the quality of Wikipedia articles," in *Proceedings of the 4th Workshop on Information Credibility*, ser. WICOW '10. New York, NY, USA: ACM, 2010, pp. 11–18. [Online]. Available: http://doi.acm.org/10.1145/1772938.1772943

[8] J. Liu and S. Ram, "Who does what: Collaboration patterns in the Wikipedia and their impact on article quality," *ACM Trans. Manage. Inf. Syst.*, vol. 2, no. 2, pp. 11:1–11:23, Jul. 2011. [Online]. Available: http://doi.acm.org/10.1145/1985347.1985352

[9] Y. Suzuki and M. Yoshikawa, "Mutual evaluation of editors and texts for assessing quality of Wikipedia articles," in *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*, ser. WikiSym '12. New York, NY, USA: ACM, 2012, pp. 18:1–18:10. [Online]. Available: http://doi.acm.org/10.1145/2462932.2462956

[10] A. Cusinato, V. Della Mea, F. Di Salvatore, and S. Mizzaro, "QuWi: Quality control in Wikipedia," in *Proceedings of the 3rd Workshop on Information Credibility on the Web*, ser. WICOW '09. New York, NY, USA: ACM, 2009, pp. 27–34. [Online]. Available: http://doi.acm.org/10.1145/1526993.1527001

[11] M. Warncke-Wang, D. Cosley, and J. Riedl, "Tell me more: An actionable quality model for Wikipedia," in *Proceedings of the 9th International Symposium on Open Collaboration*, ser. WikiSym '13.

New York, NY, USA: ACM, 2013, pp. 8:1–8:10. [Online]. Available: http://doi.acm.org/10.1145/2491055.2491063

[12] F. Peng and A. McCallum, "Information extraction from research papers using conditional random fields," *Information Processing & Management*, vol. 42, no. 4, pp. 963–979, 2006.

[13] C. Zhang, H. Wang, Y. Liu, D. Wu, Y. Liao, and B. Wang, "Automatic keyword extraction from documents using conditional random fields," *Journal of Computational Information Systems*, vol. 4, no. 3, pp. 1169–1180, 2008.

[14] K. Coursey and R. Mihalcea, "Topic identification using Wikipedia graph centrality," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, ser. NAACL-Short '09. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009, pp. 117–120. [Online]. Available: http://dl.acm.org/citation.cfm?id=1620853.1620887

[15] K. Coursey, R. Mihalcea, and W. Moen, "Using encyclopedic knowledge for automatic topic identification," in *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, ser. CoNLL '09. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009, pp. 210–218. [Online]. Available: http://dl.acm.org/citation.cfm?id=1596374.1596407

[16] Z. S. Syed, T. Finin, and A. Joshi, "Wikipedia as an ontology for describing documents," in *Proceedings of the Second International Conference on Weblogs and Social Media*. AAAI Press, 2008.

[17] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Commun. ACM*, vol. 18, no. 11, pp. 613–620, Nov. 1975. [Online]. Available: http://doi.acm.org/10.1145/361219.361220