Using Encyclopedic Knowledge for Automatic Topic Identification

Kino Coursey
University of North Texas and Daxtron Laboratories, Inc.
kino@daxtron.com

Rada Mihalcea and William Moen
University of North Texas
rada,wemoen@unt.edu

Abstract

This paper presents a method for automatic topic identification using an encyclopedic graph derived from Wikipedia. The system is found to exceed the performance of previously proposed machine learning algorithms for topic identification, with an annotation consistency comparable to human annotations.

1 Introduction

With exponentially increasing amounts of text being generated, it is important to find methods that can annotate and organize documents in meaningful ways. In addition to the content of the document itself, other relevant information about a document such as related topics can often enable a faster and more effective search or classification. Document topics have been used for a long time by librarians to improve the retrieval of a document, and to provide background or associated information for browsing by human users. They can also assist search, background information gathering and contextualization tasks, and enhanced relevancy measures.

The goal of the work described in this paper is to automatically find topics that are relevant to an input document. We refer to this task as "topic identification" (Medelyan and Witten, 2008). For instance, starting with a document on "United States in the Cold War," we want to identify relevant topics such as "history," "Global Conflicts," "Soviet Union," and so forth. We propose an unsupervised method for topic identification, based on a biased graph centrality algorithm applied to a large knowledge graph built from Wikipedia.

The task of topic identification goes beyond keyword extraction (Mihalcea and Csomai, 2007), since relevant topics may not be necessarily mentioned in the document, and instead have to be obtained from some repositories of external knowledge. The task is also different from text classification (Gabrilovich and Markovitch, 2006), since the topics are either not known in advance or are provided in the form of a controlled vocabulary with thousands of entries, and thus no classification can be performed. Instead, with topic identification, we aim to find topics (or categories[1]) that are relevant to the document at hand, which can be used to enrich the content of the document with relevant external knowledge.

2 Dynamic Ranking of Topic Relevance

Our method is based on the premise that external encyclopedic knowledge can be used to identify relevant topics for a given document.

The method consists of two main steps. In the first step, we build a knowledge graph of encyclopedic concepts based on Wikipedia, where the nodes in the graph are represented by the entities and categories that are defined in this encyclopedia. The edges between the nodes are represented by their relation of proximity inside the Wikipedia articles. The graph is built once and then it is stored offline, so that it can be efficiently use for the identification of topics in new documents.

In the second step, for each input document, we first identify the important encyclopedic concepts in the text, and thus create links between the content of the document and the external encyclopedic graph. Next, we run a biased graph centrality algorithm on the entire graph, so that all the nodes in the external knowledge repository are ranked based on their relevance to the input document. We use a variation

[1]Throughout the paper, we use the terms "topic" and "category" interchangeably.

Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL), pages 210–218,
Boulder, Colorado, June 2009.

of the PageRank (Brin and Page, 1998) algorithm, which accounts for both the relation between the nodes in the document and the encyclopedic graph, as well as the relation between the nodes in the encyclopedic graph itself.

In the following, we first describe the structure of Wikipedia, followed by a brief description of the Wikify! system that automatically identifies the encyclopedic concepts in a text, and finally a description of the dynamic ranking process on the encyclopedic graph.

2.1 Wikipedia

Wikipedia (http://wikipedia.org) is a free online encyclopedia, representing the outcome of a continuous collaborative effort of a large number of volunteer contributors. Virtually any Internet user can create or edit a Wikipedia webpage, and this "freedom of contribution" has a positive impact on both the quantity (fast-growing number of articles) and the quality (potential mistakes are quickly corrected within the collaborative environment) of this resource.

Wikipedia has grown to become one of the largest online repositories of encyclopedic knowledge, with millions of articles available for a large number of languages. In fact, Wikipedia editions are available for more than 250 languages, with a number of entries varying from a few pages to close to three million articles per language.

The basic entry in Wikipedia is an *article* (or *page*), which defines an entity or an event, and consists of a hypertext document with hyperlinks to other pages within or outside Wikipedia. The role of the hyperlinks is to guide the reader to pages that provide additional information about the entities or events mentioned in an article. Each article in Wikipedia is uniquely referenced by an identifier, which consists of one or more words separated by spaces or underscores, and occasionally a parenthetical explanation. The current version of the English Wikipedia consists of about 2.75 million articles.

In addition to articles, Wikipedia also includes a large number of categories, which represent topics that are relevant to a given article (the July 2008 version of Wikipedia includes about 390,000 such categories). The category links are organized hierarchically, and vary from broad topics such as "history" or "games" to highly focused topics such as "military history of South Africa during World War II" or "role-playing game publishing companies."

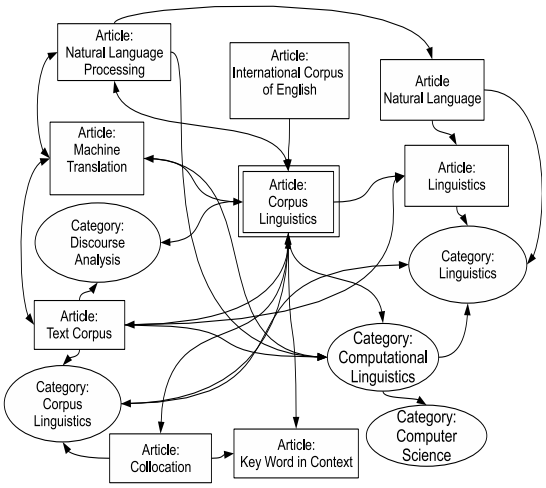

Figure 1: A snapshot from the encyclopedic graph.

We use the entire English Wikipedia to build an encyclopedic graph for use in the topic identification process. The nodes in the graph are represented by all the article and category pages in Wikipedia, and the edges between the nodes are represented by their relation of proximity inside the articles. The graph contains 5.8 million nodes, and 65.5 million edges. Figure 1 shows a small section of the knowledge graph, as built starting with the article on "Corpus Linguistics".

2.2 Wikify!

In order to automatically identify the important encyclopedic concepts in an input text, we use the unsupervised system Wikify! (Mihalcea and Csomai, 2007), which identifies the concepts in the text that are likely to be highly relevant (i.e., "keywords") for the input document, and links them to Wikipedia concepts.

Wikify! works in three steps, namely: (1) candidate extraction, (2) keyword ranking, and (3) word sense disambiguation. The candidate extraction step parses the input document and extracts all the possible n-grams that are also present in the vocabulary used in the encyclopedic graph (i.e., anchor texts for links inside Wikipedia or article or category titles).

Next, the ranking step assigns a numeric value to each candidate, reflecting the likelihood that a given candidate is a valuable keyword. Wikify! uses a "keyphraseness" measure to estimate the probability of a term W to be selected as a keyword in a

document, by counting the number of documents where the term was already selected as a keyword $count(D_{key})$ divided by the total number of documents where the term appeared $count(D_W)$. These counts are collected from all the Wikipedia articles.

$$P(keyword|W) \approx \frac{count(D_{key})}{count(D_W)} \quad (1)$$

This probability can be interpreted as "the more often a term was selected as a keyword among its total number of occurrences, the more likely it is that it will be selected again."

Finally, a simple word sense disambiguation method is applied, which identifies the most likely article in Wikipedia to which a concept should be linked to. This step is trivial for words or phrases that have only one corresponding article in Wikipedia, but it requires an explicit disambiguation step for those words or phrases that have multiple meanings (e.g., "plant") and thus multiple candidate pages to link to. The algorithm is based on statistical methods that identify the frequency of meanings in text, combined with symbolic methods that attempt to maximize the overlap between the current document and the candidate Wikipedia articles. See (Mihalcea and Csomai, 2007) for more details.

2.3 Biased Ranking of the Wikipedia Graph

Starting with the graph of encyclopedic knowledge, and knowing the nodes that belong to the input document, we want to rank all the nodes in the graph so that we obtain a score that indicates their importance relative to the given document. We can do this by using a graph-ranking algorithm *biased* toward the nodes belonging to the input document.

Graph-based ranking algorithms such as PageRank are a way of deciding the importance of a vertex within a graph, based on global information recursively drawn from the entire graph. One formulation is in terms of a random walk through a directed graph. A "random surfer" visits nodes of the graph, and has some probability of jumping to some other random node of the graph, and the remaining probability of continuing their walk from the current node to one in its outdegree list. The rank of a node is an indication of the probability that the surfer would be found at that node at any given time.

Formally, let $G = (V, E)$ be a directed graph with the set of vertices V and set of edges E, where E is a subset of $V \times V$. For a given vertex V_i, let $In(V_i)$ be the set of vertices that point to it (predecessors), and let $Out(V_i)$ be the set of vertices that vertex V_i points to (successors). The score of a vertex V_i is defined as follows (Brin and Page, 1998):

$$S(V_i) = (1-d) + d * \sum_{j \in In(V_i)} \frac{1}{|Out(V_j)|} S(V_j) \quad (2)$$

where d is a damping factor usually set to 0.85.

Given the "random surfer" interpretation of the ranking process, the $(1-d)$ portion represents the probability that a surfer will jump to a given node from any other node at random, and the summation portion indicates that the process will enter the node via edges directly connected to it.

We introduce a bias in this graph-based ranking algorithm by extending the framework of personalization of PageRank proposed by (Haveliwala, 2002). We modify the formula so that the $(1-d)$ component also accounts for the importance of the concepts found in the input document, and it is suppressed for all the nodes that are not found in the input document.

$$S(V_i) = (1-d) * Bias(V_i) + d * \sum_{j \in In(V_i)} \frac{1}{|Out(V_j)|} S(V_j) \quad (3)$$

where $Bias(V_i)$ is only defined for those nodes initially identified in the input document:

$$Bias(V_i) = \frac{f(V_i)}{\sum_{j \in InitalNodeSet} f(V_j)}$$

and 0 for all other nodes in the graph. $InitalNodeSet$ is the set of nodes belonging to the input document.

Note that $f(V_i)$ can vary in complexity from a default value of 1 to a complex knowledge-based estimation. In our implementation, we use a combination of the "keyphraseness" score assigned to the node V_i and its distance from the "Fundamental" category in Wikipedia.

The use of the $Bias$ assigned to each node means the surfer random jumps will be limited to only those nodes connected to the original query. Thus the graph-ranking process becomes biased and focused on those topics directly related to the input. It also accumulates activation at those nodes not directly found in the input text, but linked through indirect means, thus reinforcing the nodes where patterns of activation intersect and creating a constructive interference pattern in the network. These reinforced nodes are the "implied related topics" of the text.

3 Illustration

To illustrate the ranking process, consider as an example the following sentence "The United States was involved in the Cold War."

First the text is passed through the Wikify! system, which returns the articles "United States" and "Cold War." Taking into account their "keyphraseness" as calculated by Wikify!, the selections are given an initial bias of 0.5492 ("United States") and 0.4508 ("Cold War").

After the first iteration the initial activation spreads out into the encyclopedic graph, the nodes find a direct connection to one another, and correspondingly their scores are changed to 0.3786 ("United States") and 0.3107 ("Cold War"). After the second iteration, new nodes are identified from the encyclopedic graph, a subset of which is shown in Figure2. The process will eventually continue for several iterations until the scores of the nodes do not change. The nodes with the highest scores in the final graph are considered to be the most closely related to the input sentence, and thus selected as relevant topics.

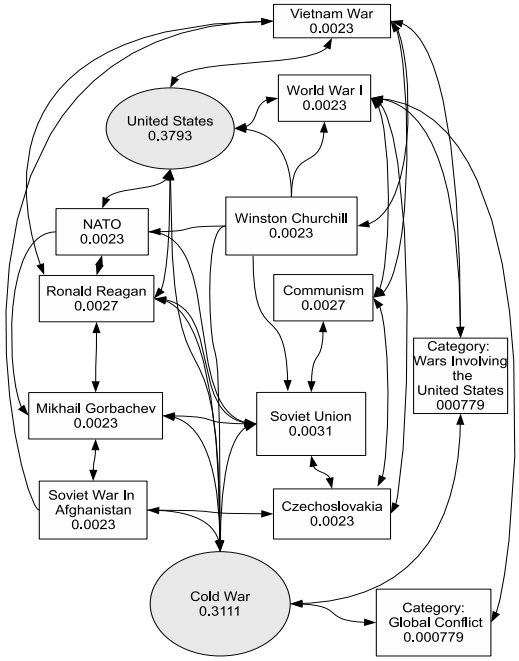

Figure 2: Sub-graph between "United States" and "Cold War"

In order to see the effect of the initial bias, consider as an example the ranking of the nodes in the encyclopedic graph when biased with the sentence "The United States was involved in the Cold War," versus the sentence "Microsoft applies Computer Science." A comparison between the scores of the nodes when activated by each of these sentences is shown in Table 1.

Wikipedia entry	US/CW	MS/CS	Diff.
A: United States	0.393636	0.006578	0.387058
C: Computer Science	0.000004	0.003576	-0.003571
A: World War II	0.007102	0.003674	0.003428
A: United Kingdom	0.005346	0.002670	0.002676
C: Microsoft	0.000001	0.001839	-0.001837
C: Cold War	0.001695	0.000006	0.001689
C: Living People	0.000835	0.002223	-0.001387
C: Mathematics	0.000029	0.001337	-0.001307
C: Computing	0.000008	0.001289	-0.001280
C: Computer Pioneers	0.000002	0.001238	-0.001235

Table 1: Node ranking differences when the encyclopedic graph is biased with different inputs: (1) "United States" and "Cold War" (US/CW) vs. (2) "Microsoft" and "Computer Science" (MS/CS). The nodes are either article pages (A) or category pages (C).

4 Experiments

In order to measure the effectiveness of the topic ranking process, we run three sets of experiments, aimed at measuring the relevancy of the automatically identified topics with respect to manually annotated gold standard data sets.

In the first experiment, the identification of the important concepts in the input text (used to bias the topic ranking process) is performed manually, by the Wikipedia users. In the second and third experiment, the identification of these important concepts is done automatically, by the Wikify! system. In all the experiments, the ranking of the concepts from the encyclopedic graph is done automatically by using the dynamic ranking process described in Section 2.

In the first two experiments, we use a data set consisting of 150 articles from Wikipedia, which have been explicitly removed from the encyclopedic graph. All the articles in this data set include manual annotations of the relevant categories, as assigned by the Wikipedia users, against which we can measure the quality of the automatic topic assignments. The 150 articles have been randomly selected while following the constraint that they each contain at least three article links and at least three category links. Our task is to rediscover the relevant categories for each page. Note that the task is nontrivial, since there are approximately 390,000 categories to choose from. We evaluate the quality of our system through the standard measures of preci-

sion and recall.

4.1 Manual Annotation of the Input Text

In this first experiment, the articles in the gold standard data set also include manual annotations of the important concepts in the text, i.e., the links to other Wikipedia articles as created by the Wikipedia users. Thus, in this experiment we only measure the accuracy of the dynamic topic ranking process, without interference from the Wikify! system.

There are two main parameters that can be set during a system run. First, the set of initial nodes used as bias in the ranking can include: (1) the initial set of articles linked to by the original document (via the Wikipedia links); (2) the categories listed in the articles linked to by the original document[2]; and (3) both. Second, the dynamic ranking process can be run through propagation on an encyclopedic graph that includes (1) all the articles from Wikipedia; (2) all the categories from Wikipedia; or (3) all the articles and the categories from Wikipedia.

Figures 3, 4 and 5 show the precision, recall and F-measure obtained for the various settings. In the plots, $Bias$ and $Propagate$ indicate the selections made for the two parameters, which can be each set to $Articles$, $Categories$, or $Both$. Each of these correspond to the options listed before.

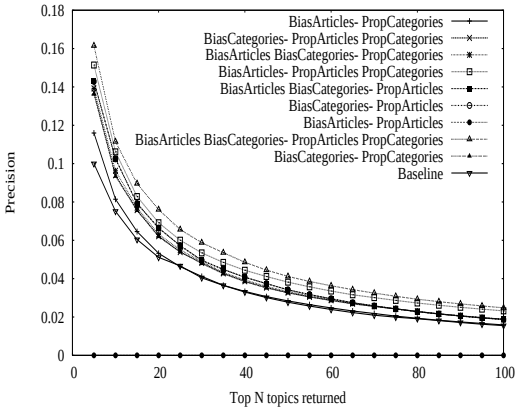

Figure 3: Precision for manual input text annotations.

As seen in the figures, the best results are obtained for a setting where both the initial bias and the propagation include all the available nodes, i.e., both articles and categories. Although the primary task is the identification of the categories, the addition of the article links improves the system performance.

[2]These should not be confused with the categories included in the document itself, which represent the gold standard annotations and are not used at any point.

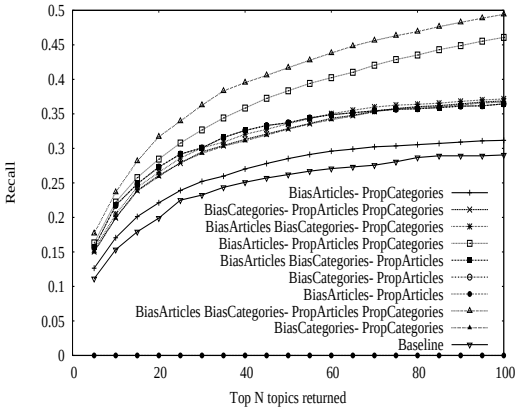

Figure 4: Recall for manual input text annotations.

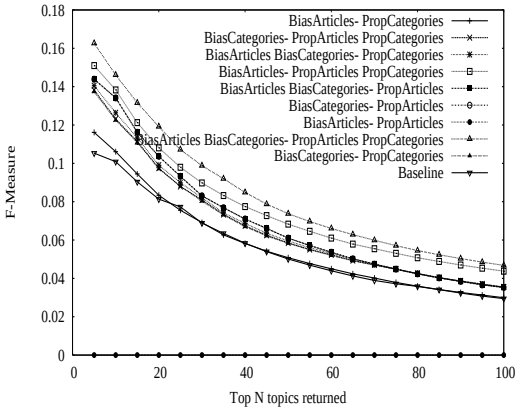

Figure 5: F-measure when using Wikipedia article annotations.

To place results in perspective, we also calculate a baseline (labeled as "Baseline" in the plots), which selects by default all the categories listed in the articles linked to by the original document. Each baseline article assigns $1/N$ to each of its N possible categories, with categories pointed to by multiple articles receiving the summation.

4.2 Automatic Annotation of the Input Text

The second experiment is similar to the first one, except that rather than using the manual annotations of the important concepts in the input document, we use instead the Wikify! system that automatically identifies these important concepts by using the method briefly described in Section 2.2. The article links identified by Wikify! are treated in the same way as the human anchor annotations from the previous experiment. In this experiment, we have

an additional parameter, which consists of the percentage of links selected by Wikify! out of the total number of words in the document. We refer to this parameter as keyRatio. The higher the keyRatio, the more terms are added, but also the higher the potential of noise due to mis-disambiguation.

Figures 6, 7 and 8 show the effect of varying the value of the keyRatio parameter used by Wikify! has on the precision, recall and F-measure of the system. Note that in this experiment, we only use the best setting for the other two parameters as identified in the previous experiment, namely an initial bias and a propagation step that include all available nodes, i.e., both articles and categories.

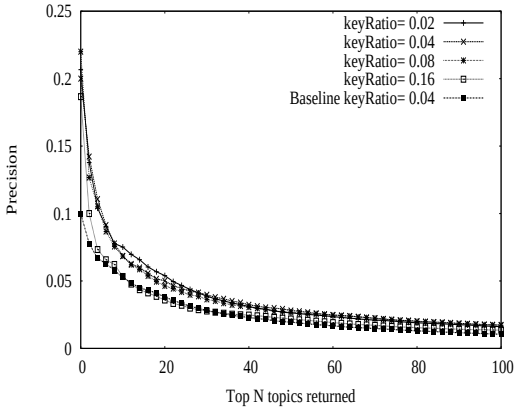

Figure 6: Precision for automatic input text annotations (Wikipedia data set)

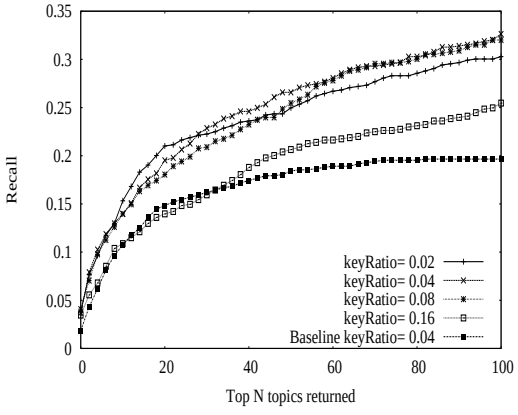

Figure 7: Recall for automatic input text annotations (Wikipedia data set)

The system's best performance occurs for a keyRatio of 0.04 to 0.06, which coincides with the ratio found optimal in previous experiments using the Wikify! system (Mihalcea and Csomai, 2007).

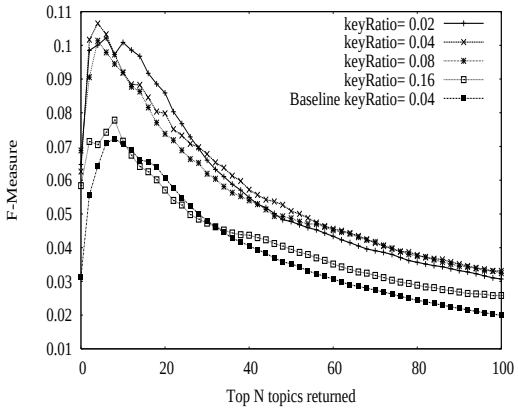

Figure 8: F-measure for automatic input text annotations (Wikipedia data set)

As before, we also calculate a baseline, which selects by default all the categories listed in the articles linked to by the original document, with the links being automatically identified with the Wikify! system. The baseline is calculated for a keyRatio of 0.04, which is one of the values that were found to work well for the ranking system itself and in previous Wikify! experiments.

Overall, the system manages to find many relevant topics for the documents in the evaluation data set, despite the large number of candidate topics (close to 390,000). Our system exceeds the baseline by a large margin, demonstrating the usefulness of using the biased ranking on the encyclopedic graph.

4.3 Article Selection for Computer Science Texts

In the third experiment, we use again the Wikify! system to annotate the input documents, but this time we run the evaluations on a data set consisting of computer science documents. We use the data set introduced in previous work on topic identification (Medelyan and Witten, 2008), where 20 documents in the field of computer science were independently annotated by 15 teams of two computer science undergraduates. The teams were asked to read the texts and assign to each of them the title of the five Wikipedia articles they thought were the most relevant **and** the other groups would also select. Thus, the consistency of the annotations was an important measure for this data set. (Medelyan and Witten, 2008) define consistency as a measure of agreement:

$$Consistency = \frac{2C}{A+B}$$

where A and B are the number of terms assigned by two indexing teams, and C is the number of terms they have in common. In the annotations experiments reported in (Medelyan and Witten, 2008), the human teams consistency ranged from 21.4% to 37.1%, with 30.5% being the average.[3]

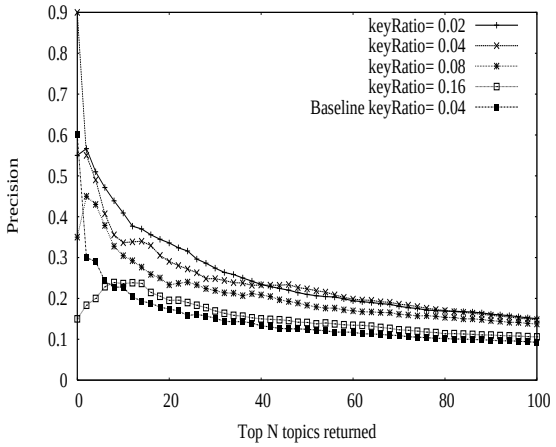

Figure 9: Precision for automatic input text annotations (Waikato data set)

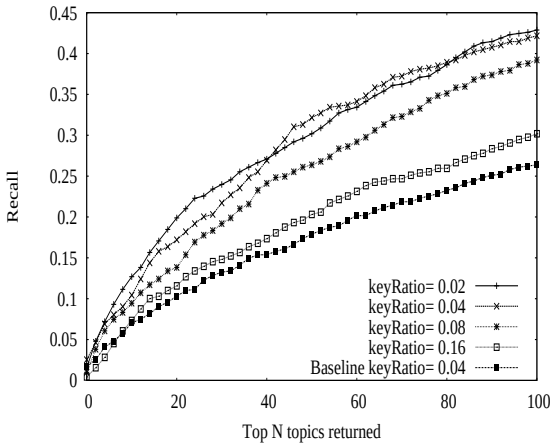

Figure 10: Recall for automatic input text annotations (Waikato data set)

Figures 10, 9, 11 and 12 show the performance of our system on this data set, by using the Wikify! annotations for the initial bias, and then propagating to both articles and categories. The plots also show a baseline that selects all the articles automatically identified in the original document by using the Wikify! system with a keyRatio set to 0.04.

[3]The consistency for one team is measured as the average of the consistencies with the remaining 14 teams.

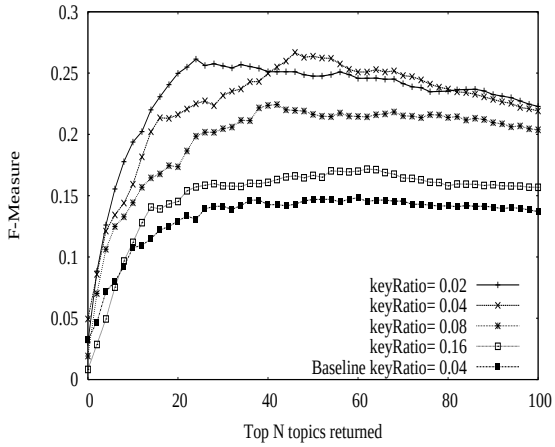

Figure 11: F-measure for automatic input text annotations (Waikato data set)

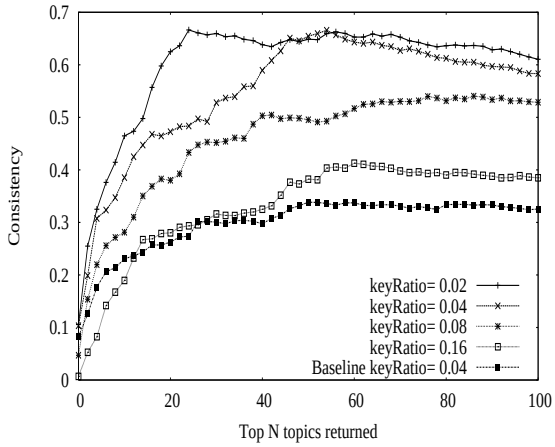

Figure 12: Consistency for automatic input text annotations (Waikato data set)

When selecting the top five topics returned by our system (the same number of topics as provided by the human teams), the average consistency with respect to the 15 human teams was measured at 34.5%, placing it between the 86% and 93% percentile of the human participants, with only two human teams doing better. We can compare this result with the one reported in previous work for the same data set. Using a machine learning system, (Medelyan and Witten, 2008) reported a consistency of 30.5%. Thus, our result of 34.5% is significantly better, despite the fact that our method is unsupervised.

In a second evaluation, we also considered the union of all the terms assigned by the 15 teams. On average, each document was assigned 35.5 different terms by the human teams. If allowed to provide more annotations, our system peaks with a con-

sistency of 66.6% for the top 25 topics returned. The system has the ability to identify possible relevant alternative topics using the comprehensive catalog of Wikipedia computer science articles and their possible associations. A human team may not necessarily consider all of the possibilities or even be aware that some of the articles, possibly known and used by the other teams, exist.

5 Related Work

The work closest to ours is perhaps the one described in (Medelyan and Witten, 2008), where topics relevant to a given document are automatically selected by using a machine learning system. Unlike our unsupervised approach, (Medelyan and Witten, 2008) learn what makes a good topic by training on previously annotated data.

Also related is the Wikify! system concerned with the automatic annotation of documents with Wikipedia links (Mihalcea and Csomai, 2007). However, Wikify! is purely extractive, and thus it cannot identify important topics or articles that are not explicitly mentioned in the input text.

Explicit semantic analysis (Gabrilovich and Markovitch, 2006) was also introduced as a way to determine the relevancy of the Wikipedia articles with respect to a given input text. The resulting vector however is extremely large, and while it was found useful for the task of text classification with a relatively small number of categories, it would be difficult to adapt for topic identification when the number of possible topics grows beyond the approximately 390,000 under consideration. In a similar line of work, (Bodo et al., 2007) examined the use of Wikipedia and latent semantic analysis for the purposes of text categorization, but reported negative results when used for the categorization of the Reuters-21578 dataset.

Others are exploring the use of graph propagation for deriving semantic information. (Hughes and Ramage, 2007) described the use of a biased PageRank over the WordNet graph to compute word pair semantic relatedness using the divergence of the probability values over the graph created by each word. (Ollivier and Senellart, 2007) describes a method to determine related Wikipedia article using a Markov chain derived value called the green measure. Differences exist between the PageRank based methods used as a baseline in their work and the method proposed here, since our system can use the content of the article, multiple starting points, and tighter control of the random jump probability via the bias value. Finally, (Syed et al., 2008) reported positive results by using various methods for topic prediction including the use of text similarity and spreading activation. The method was tested by using randomly selected Wikipedia articles, where in addition to the categories listed on a Wikipedia page, nearby subsuming categories were also included as acceptable.

6 Conclusions and Future Work

In this paper, we introduced a system for automatic topic identification, which relies on a biased graph centrality algorithm applied on a richly interconnected encyclopedic graph built from Wikipedia. Experiments showed that the integration of encyclopedic knowledge consistently adds useful information when compared to baselines that rely exclusively on the text at hand. In particular, when tested on a data set consisting of documents manually annotated with categories by Wikipedia users, the topics identified by our system were found useful as compared to the manual annotations. Moreover, in a second evaluation on a computer science data set, the system exceeded the performance of previously proposed machine learning algorithms, which is remarkable given the fact that our system is unsupervised. In terms of consistency with manual annotations, our system's performance was found to be comparable to human annotations, with only two out of 15 teams scoring better than our system.

The system provides a means to generate a dynamic ranking of topics in Wikipedia within a framework that has the potential to utilize knowledge or heuristics through additional resources (like ontologies) converted to graph form. This capability is not present in resources like search engines that provide access to a static ranking of Wikipedia. Future work will examine the integration of additional knowledge sources and the application of the method for metadata document annotations.

Acknowledgments

This work has been partially supported by an award #CR72105 from the Texas Higher Education Coordinating Board and by an award from Google Inc. The authors are grateful to the Waikato group for making their data set available.

References

Z. Bodo, Z. Minier, and L. Csato. 2007. Text categorization experiments using Wikipedia. In *Proceedings of the International Conference on Knowledge Engineering ,Principles and Techniques*, Cluj-Napoca (Romania).

S. Brin and L. Page. 1998. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7).

E. Gabrilovich and S. Markovitch. 2006. Overcoming the brittleness bottleneck using Wikipedia: Enhancing text categorization with encyclopedic knowledge. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, Boston.

T. Haveliwala. 2002. Topic-sensitive PageRank. In *Proceedings of the Eleventh International World Wide Web Conference*, May.

T. Hughes and D. Ramage. 2007. Lexical semantic knowledge with random graph walks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Prague, Czech Republic.

O. Medelyan and I. H. Witten. 2008. Topic indexing with Wikipedia. In *Proceedings of the AAAI WikiAI workshop*.

R. Mihalcea and A. Csomai. 2007. Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management*, Lisbon, Portugal.

Y. Ollivier and P. Senellart. 2007. Finding related pages using green measures: An illustration with wikipedia. In *Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence (AAAI 2007)*.

Z. Syed, T. Finin, and A. Joshi. 2008. Wikipedia as an Ontology for Describing Documents. In *Proceedings of the Second International Conference on Weblogs and Social Media*. AAAI Press, March.