# Topic Identification Using Wikipedia Graph Centrality

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

**Rada Mihalcea**
University of North Texas
rada@cs.unt.edu

## Abstract

This paper presents a method for automatic topic identification using a graph-centrality algorithm applied to an encyclopedic graph derived from Wikipedia. When tested on a data set with manually assigned topics, the system is found to significantly improve over a simpler baseline that does not make use of the external encyclopedic knowledge.

## 1 Introduction

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 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, 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, 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.

### 2.1 Wikipedia

Wikipedia (http://en.wikipedia.org) is a free online encyclopedia, representing the outcome of a continuous collaborative effort of a large number of volunteer contributors. The basic entry is an *article*, which defines an entity or an event, and consists of a hypertext document with hyperlinks to other pages within or outside Wikipedia. In addition to arti-

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



*Proceedings of NAACL HLT 2009: Short Papers*, pages 117–120,
Boulder, Colorado, June 2009. 

cles, 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 more than 350,000 such categories).

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.

### 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 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.

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

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. 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 essentially 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. The rank of a node is an indication of the probability that one would find the surfer 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 PageRank score of a vertex $V_i$ is defined as follows (Brin and Page, 1998):

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

where d is a damping factor usually set to 0.85.

In a "random surfer" interpretation of the ranking process, the $(1-d)$ portion represents the probability that a surfer navigating the graph 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. Using a method inspired by earlier work (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.

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

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.

## 3 Experiments

We run two experiments, aimed at measuring the relevancy of the automatically identified topics with respect to a manually annotated gold standard data set.

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 experiment, the identification of these important concepts is done automatically with the Wikify! system. In both experiments, the ranking of the concepts from the encyclopedic graph is performed using the dynamic ranking process described in Section 2.

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 non-trivial, since there are more than 350,000 categories to choose from. We evaluate the quality of our system through the standard measures of precision and recall.

### 3.1 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 1 and 2 show the precision and recall for the various settings. $Bias$ and $Propagate$ indicate the selections made for the two parameters, which can be set to either $Articles$, $Categories$, or $Both$.

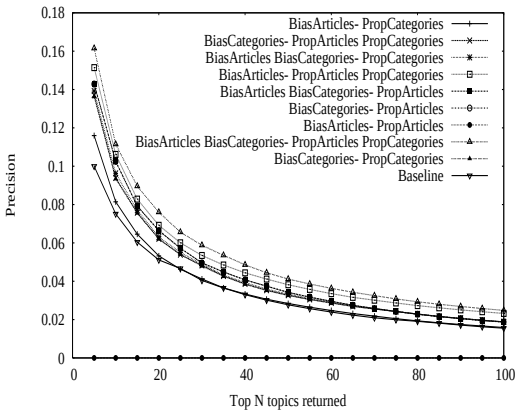


Figure 1: Precision for manual input text annotations.

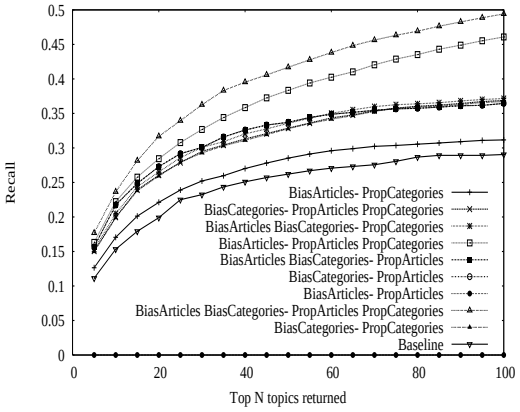


Figure 2: Recall for manual input text annotations.

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.

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.

### 3.2 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 3 and 4 show the effect of varying the value of the keyRatio parameter on the precision and recall 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.

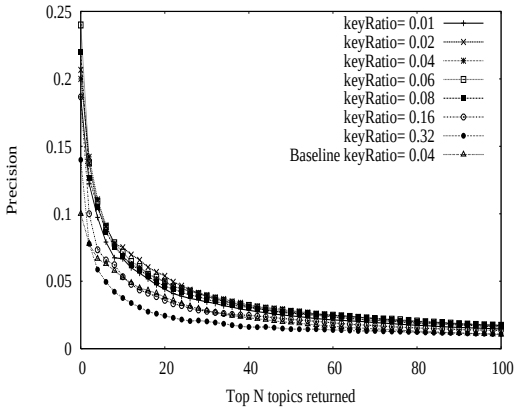


Figure 3: Precision for automatic input text annotations

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

Overall, the system manages to find many relevant topics for the documents in the evaluation data set, despite the large number of candidate topics (more than 350,000). Additional experiments performed against a set of documents from a source other than Wikipedia are reported in (Coursey et al., 2009).

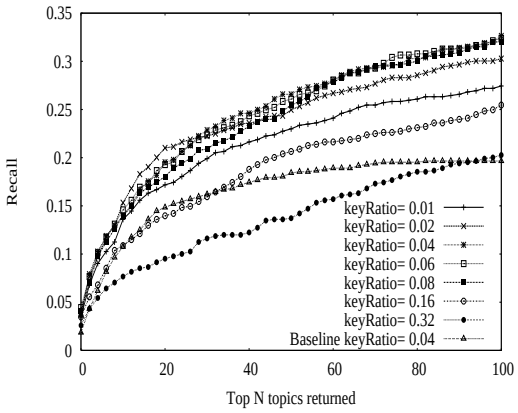


Figure 4: Recall for automatic input text annotations

## 4 Conclusions

In this paper, we presented an unsupervised system for automatic topic identification, which relies on a biased graph centrality algorithm applied on a graph built from Wikipedia. Our experiments demonstrate the usefulness of external encyclopedic knowledge for the task of topic identification.

## Acknowledgments

This work has been partially supported by 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

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

K. Coursey, R. Mihalcea, and W. Moen. 2009. Using encyclopedic knowledge for automatic topic identification. In *Proceedings of the Conference on Natural Language Learning*, Boulder, CO.

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

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.