# Discovering Generalized Concepts from Documents Using a Category Graph

Tom Vacek
Dept. of Computer Science
University of Minnesota
Minneapolis, MN 55455
vacek@cs.umn.edu

John Joseph
Dept. of Computer Science
University Of Minnesota
Minneapolis, MN 55455
jjoseph@cs.umn.edu

Daniel Boley
Dept. of Computer Science
University Of Minnesota
Minneapolis, MN 55455
boley@cs.umn.edu

## Abstract

Most concept assignment methods assign concepts closest to each specific document in some sense, while users might be interested in a somewhat broader topic area. For example, topics for a document on the San Jose Sharks could be San Jose Sharks, Hockey Team Mascots, NHL Teams, National Hockey League, Hockey, or Sports, depending on the context. Many papers have propsed techniques to use Wikipedia to assign topics to standalone documents. This paper proposes a follow-on technique to find another set of concept tags that are more general. The technique is a modification of the page rank algorithm and takes a user-selectable parameter to bias a random walk of the Wikipedia category graph toward the roots or the leaves. Unlike other extant techniques this parameter allows users to change the generality or specificity of the discovered concepts. This also facilitates finding common topic themes in a multiple-document set. Applications for 'the proposed technique include automating the organization and search of dynamic repositories such as newsgroups, email, text messages, Tweets, and the like.

## 1 Introduction

Topic indexing has traditionally been a major part of information retrieval. To this day, library cataloguing systems are organized around an encyclopedic division of knowledge, such as the Library of Congress or Dewey systems. New items to be catalogued are assigned their topics by human indexers, who also maintain the hierarchy of topics. Though keyword-based information retrieval techniques have given researchers alternatives to using topics to search for information, topics still have many uses in information retrieval. For instance, consult [13] for a discussion of a scatter-gather search interface.

There are several approaches to automated topic indexing, and the following brief overview adopts the terminology from [16]. Keyphrase extraction uses statistical techniques to extract short phrases or n-grams from the document itself. Term assignment uses a classifier trained on sample documents from a human-organized list of topics to make new topic assignments. Finally, keyphrase indexing is a hybrid of these two, first extracting significant phrases from the document and then using external knowledge to map them into a controlled vocabulary.

For either of the last two approaches, significant human labor is required. For term assignment, a topic ontology along with a set of training documents for each topic must be laboriously curated. For any large ontology, changes in understanding and the advent of new knowledge invariably force modification, so curating an ontology requires significant work by experts. Likewise, keyphrase indexing requires a vocabulary and a corpus of external knowledge from which semantic information can be drawn.

Even if automated topic assignment is just as good as manual classification, the static nature of the topics assigned in both of these approaches presents another problem. For example, one person might think that *token-ring protocols* is the best topic for some document, while another person would want *computer*

*networks*. This example shows that the perspective of the information seeker is a factor in what a topic should be for a document. In order for a topic assignment system to produce topics useful to a user, the context of the topic ontology needs to match the perspective of the user.

A number of papers have proposed using Wikipedia in term assignment and keyphrase indexing. Wikipedia is a convenient source of external knowledge for these applications since it has become one of the largest repositories of human knowledge. Moreover, Wikipedia's structure—a large number of articles that tagged with hierarchical categories—is remarkably well-suited for these applications.

The distributed nature of Wikipedia's development is a double-edged sword. On one hand, it has allowed phenomenal growth. On the other hand, the articles and categories do not have the uniform quality that one would find a professionally-maintained ontology. Wikipedia articles are less developed in some knowledge areas than others. In the category ontology, approaches to what constitutes a category vary considerably. Some topics are 'nominalistic,' that is, the topic encompasses objects that have some nonessential quality in common, such as *1891 establishments*. Other topics are very specific, such as *New York City subway passenger equipment* or *Osaka municipal subway stations*, and appear in a part of the category graph that is far more developed than other parts. Documents with such a topic occur so infrequently that any classifier will face a severe asymmetry problem. Moreover, a person might not find such a topic useful. Perhaps *public transportation* or *railroad equipment* would be more useful.

In this paper we present a technique to use information contained the structure of the Wikipedia category graph to try to overcome these problems. If a document is tagged with topics from Wikipedia's categories, this technique returns a new set of tags, also from Wikipedia's categories. The goal of our algorithm is threefold: first, to avoid nominalistic categories as much as possible, second, to provide robustness against incorrect taggings in the initial tagging, and finally, to assign topics from a perspective which is useful to the user. For the last objective, our algorithm has a user-selectable parameter which provides some control over the generality of the returned topics.

For this paper, we implemented a topic assignment technique which first uses tf-idf term assignment to score the relevance of each Wikipedia category for a given document. We call these preliminary topics "base concepts." Then we apply our technique and call the results "generalized topics." Our base concepts implementation is intended primarly as a platfrom for the second-step technique and could be replaced by more sophisticated techniques, such as ones used in [16], [9], or [5].

Section 1.1 presents the related work in this area. Section 1.2 gives a short summary of the structure of Wikipedia ontology used in our work. In section 1.3 we discuss the base concepts, the properties they should satisfy, and the simple method we chose to obtain them. In section 2 we describe our proposed technique to find generalized comments from the initial set of base concepts. In section 3 we evaluate the effectiveness our technique using two approaches. In section 4 we give some concluding remarks and future work, including the possible use of more sophisticated base concepts.

### 1.1 Background and Previous work

The concept identification problem has been traditionally studied by the machine learning and natural language processing community under titles such as topic extraction or key phrase extraction. Many of these techniques focus on concept discovery on a collection of documents [10] as against finding topics on isolated documents. Traditional text clustering techniques such as the k-means algorithm addresses part of the issues related to concept discovery on collections of documents, in the sense that they group together related documents.

For isolated documents, most topic discovery techniques [11, 23, 7, 15] use keyphrase extraction. This approach is sufficient to associate a document or cluster with a concept, however, term assignment and keyphrase indexing promise better results. For example, it is possible to write an entire article that discusses *moon landing* without ever mentioning the

words *space exploration*. Keyphrase extraction will not be able to associate a broader concept such as *space exploration* to this document. The performance of n-grams to help classify the genre of documents was studied in [14].

Techniques exist for concept identification that place documents in an external ontology of concepts such as WordNet [24]. WordNet groups words describing a related concept into synsets. Synsets are linked by semantic relationships such as hypernyms and hyponyms, that indicate class-subclass relationship between these synsets. The biggest drawback of this approach is that the WordNet ontology is hand constructed by a few lexicographic experts and these synsets cover only a fraction of the entire possible concepts. Moreover, synsets can be thought of more as a group of synonyms rather than actual concepts that capture world knowledge. Hierarchical directories of web sites such as Yahoo Directories or Open Directory Project have also been used for concept identification [22, 12].

Considerable research has been done to utilize Wikipedia to identify semantic relatedness among documents or among words. [8] shows that a method much like our base concepts procedure can be used to compare the semantic relatedness of two documents, and [9] builds upon these results. [20] attempts to find semantic relatedness at the word level. Nastase and Strube [18] describe techniques by which word-level semantic relationships can be inferred from the Wikipedia category information.

Wikipedia has been used to improve keyword extraction. Wikify! [17] extracts keywods from documents and applies word-sense disambiguation to link these keywords to the appropriate Wikipedia articles, as opposed to categories, as done here.

A number of articles propose to use Wikipedia for automated topic indexing. [16] uses Wikipedia for keyphrase indexing, where the controlled vocabulary is made up of the titles of Wikipedia articles. [19] proposes a way to rank Wikipedia categories by their relevance to a document, where a combination of term assignment and keyphrase extraction methods are used.

The articles most similar to ours also try to generalize the topics assigned in some preliminary step. First, [21] uses cosine similarity, like our base concepts algorithm, to find the Wikipedia articles most relevant to an input document, and then it finds Wikipedia categories based on those articles. Finally, they use the spreading activation algorithm on the category graph to generalize the results. It is difficult to compare results since they evaluate their results using articles removed from Wikipedia, which still share features in common with the remaining portion of Wikipedia, from which similar articles are found. Thus, it is an easier dataset than ours. Nevertheless, it is apparent that spreading activation allows only one or two possibilities for the generality of returned topics.

Finally, [5, 4] make use of the Wikify! [17] system to identify the Wikipedia articles that are most relevant to the document and then use a biased PageRank algorithm on a graph made up of both article links and categories. The random walk is biased toward the articles identified in the first step. The method ignores the hierarchical information found in Wikipedia's category graph, treating it in an undirected fashion. This portion of their work is not evaluated in a way comparable to ours.

We limit our method to the Wikipedia category graph, taking advantage of the fact that this is a directed graph, unlike the the article links. Thus, we use the category graph structure to provide mechanisms to control the generality of the discovered concepts as needed. Moreover, [3] suggests that the information contained in the graph structure of the article links is not the same as the information in the category ontology, so we believe conflating the two requires further justification.

Finally, [1] and [2] present methods to match two different hierarchical structures. Our paper makes use of a hierarchical ontology to identify a topic for a standalone document, but once a topic has been identified we do not consider its place in the hierarchy to be important. Therefore, these two papers have limited application to our project.

## 1.2 Category Ontology

We use the Wikipedia dataset as provided by the INXS 2007 Workshop [6] for our experiments. The dataset

contains 659,388 different articles in 115,625 different categories. Although this dataset contains only a subset of articles and categories in Wikipedia, we found that this dataset is sufficient for the experiments in this paper. Note that the present Wikipedia category graph is more extensive with more than 390,000 categories[5].

We did a small amount of preprocessing. Motivated by our goal to avoid nominalistic topics, we simply removed the offenders that we could easily find from Wikipedia's category graph. Examples of categories removed include *1951 births* or *february 5 deaths*. Nevertheless, a great number remain, such as *conflicts in 1941*, so it is necessary for the technique to provide some robustness against them. As a practical matter, we removed all Wikipedia categories that have neither a parent nor a child category. After these steps, the resulting categories form a directed graph of about 80,000 categories. The graph does contain cycles, likely the result of the distributed nature of Wikipedia's development.

### 1.3 Base Similarity

The first step of this procedure is to identify the base concepts for a document. and then an independent algorithm generalizes them. We mentioned that a number of techniques would do. The base concepts procedure ought to satisfy the following axioms: First, the algorithm scores the relevance of every element in the category ontology to a given document (though these scores could be almost entirely zero). Second, concepts identified ought to be topics of some portion of the document in some sense. (For instance, a document about graphics cards should have a much higher score for "graphics cards" than "playing cards.") Our procedure uses cosine similarity between a given document and all the documents in a Wikipedia category. By repeating this for every category in Wikipedia, we satisfy the first axiom. Our approach, however, using could be improved with respect to the second axiom.

#### Calculating base scores

Let $C = \{C_1, C_2, ..., C_i, ...C_N\}$ be the set of all the Wikipedia categories. Let $D_i$ be the set of Wikipedia articles that are assigned to the category $C_i$. We let $s_i^d$ be the base similarity of the given document $d$ with the category $C_i$. If $|D_i| \geq 10$, we calculate $s_i^d$ as the average of the cosine similarity of each $d_i$ with $d$:

$$s_i^d = \frac{1}{|D_i|} \sum_{d_j \in D_i} \frac{d_j^T d}{\|d_j\|_2 \|d\|_2}$$

(where $d_k$ is the tf-idf vector for that document[1]). Otherwise, $s_i^d$ is 'smoothed' to the mean of neighboring nodes.

Our vector space corresponds to a dictionary that does not include every word in the Wikipedia corpus. We used a stemmed concatenation of several common dictionaries, containing about 80,000 words.

Table 2 shows the categories with the highest base scores for the test documents in Table 1.

## 2 Generalization

The second and primary step of the proposed technique involves performing a random walk over the Wikipedia category graph in such a way that, with a high probability, a step is taken towards a node that has a high cosine score. The final, 'generalized scores' are then the probabilities associated with the stationary distribution of the random walk. At every node $i$,

- take a step towards a parent category $u$ of $i$ with a probability that is proportional to $\beta s_u^d$. If $i$ has no parent categories, then teleport to a random category $j$ with probability proportional to $s_j^d$.
- take a step towards a child category $v$ of $i$ with a probability that is proportional to $(1-\beta)s_v^d$. If $i$ has no children, teleport to a random category as above.
- teleport to a random category $j$ with probability $(1-\alpha)s_j^d$

[1]The idf factors are calculated with respect to the entire corpus of Wikipedia articles. The same idf corrections are also applied to the term frequency vector of $d$, the document for which we are finding topics.

| D1 : *rec.sport.hockey/54117* |
|---|
| From: rick@emma.tfbbs.wimsey.bc.ca (Rick Younie)<br>Subject: stats for hockey pool<br>I'm the keeper of the stats for a family hockey pool and I'm looking for daily/weekly email servers for playoff stats. I've connected with the servers at J.Militzok@skidmore.EDU and wilson@cs.ucf.edu. I'm still sorting these two out. Are there others? Email please as my site doesn't get this group. Thanks. Rick – rick@emma.panam.wimsey.bc.ca rick@emma.tfbbs.wimsey.bc.ca |
| D2 : *rec.autos/101577* |
| From: kenyon@xqzmoi.enet.dec.com (Doug Kenyon (Stardog Champion))<br>Subject: Re: Integra GSR (really about other cars)<br>It's great that all these other cars can out-handle, out-corner, and out-accelerate an Integra.<br>But, you've got to ask yourself one question: do all these other cars have a moonroof with a sliding sunshade? No wimpy pop-up sunroofs or power sliding roofs that are opaque. A moonroof that can be opened to the air, closed to let just light in, or shaded so that nothing comes in.<br>You've just got to know what's important.<br>–Doug '93 Integra GS |
| D3 : *comp.os.ms-windows.misc/9570* |
| From: narlochn@kirk.msoe.edu<br>Subject: last<br>I have two questions: 1) I have been having troubles with my Wordperfect for Windows. When I try to select and change fonts, etc. some of the text disappears. I tried to center two lines once, and the second line disappeared. I can not find the error, and I do not know how to correct it. 2) Is this the right newsgroup? Where should I go? E-mail prefered...<br>Who else is still waiting for "Naked Gun Part (Pi)" |
| D4 : *sci.med/58046* |
| From: Lawrence Curcio <lc2b+@andrew.cmu.edu><br>Subject: Analgesics with Diuretics<br>I sometimes see OTC preparations for muscle aches/back aches that combine aspirin with a diuretic. The idea seems to be to reduce inflammation by getting rid of fluid. Does this actually work?<br>Thanks, –Larry C. |
| D5 : *soc.religion.christian/20501* |
| From: harwood@umiacs.umd.edu (David Harwood)<br>Subject: Re: Essene New Testament<br>[William Christie asked about the Essene NT. Andrew Kille reponded There is a collection of gospels which usually goes under the name of the "Essene Gospel of Peace." These are derived from the gnostics, not the essenes, and are ostensibly translations from syriac texts of the fourth and fifth centuries (I vaguely recall; I can't find my copy right now). –clh]<br>There had been recent criticism of this in a listserv for academic Biblical scholars: they all say the book(s) are modern fakes. D.H. |

Table 1: This table shows five documents selected from the 20 Newsgroups data set to qualitatively demonstrate the effectiveness our method.

Here $\beta$, $0 \leq \beta \leq 1$, is an adjustable bias toward top-level categories over bottom-level categories, where $\beta = 0.5$ is neutral, and $\alpha$ is the teleportation constant as used in PageRank and is set to 0.85.

In matrix form, create two matrices $F$ and $R$ such that $F_{ij} = s_j^d$ if $j$ is the parent of $i$ in the category graph and $R_{ij} = s_j^d$ if $i$ is the parent of $j$. $F$ and $R$ are then row normalized, with zero rows replaced by $s$, the normalized vector of similarity values $s_i^d$. The stochastic matrix $P$ is given by

$$P = \alpha(\beta F + (1-\beta)R) + (1-\alpha)e\hat{s}'$$

where $e$ is the vector of all ones and $\hat{s}$ is a sparsified $s$. The stationary distribution containing the generalized scores is the left eigenvector of $P$.

# 3 Evaluation

Evaluating the success of the concept generalization technique is difficult due to the absence of a data set that is tagged with the most relevant Wikipedia categories. One approach followed in [19, 5, 4] involves removing a subset of Wikipedia articles from the Wikipedia data set as a 'test set' and then evaluating the classification performance on this test set using the original categories as labels. In our case, such a subset of Wikipedia would make a poor test set, since our goal is to find general concepts, but Wikipedia's guidelines encourage authors to tag articles with the most specific applicable categories. Rather, we use a dataset external to Wikipedia that consists in sets of documents with a similar topic, and show that the method is able to identify the unifying topic for a set. We also select a few documents to show how it works for a particular document.

We use the well-known 20 Newsgroups data set for our evaluation. It is a collection of approximately 20,000 newsgroup documents, split across 20 different newsgroups. Each of the newsgroups are themed to discuss topics under a certain domain (like *sci.space*). But some newsgroups (like *misc.forsale* and *talk.politics.misc*) are very general and can discuss a wide range of issues. Some of the newsgroups

are also very closely related to each other, such as *comp.sys.ibm.pc.hardware* and *comp.sys.mac.hardware*.

### 3.1 Qualitative evaluation: Individual documents

For the documents in Table 1, the first column of Table 2 shows the top ranking categories when sorted simply based on their base similarity score. It can be seen that a few of the top categories are related to the document but some are irrelevant. For example, for the document D1, both *server*- and *hockey*-related concepts are present, along with some irrelevant categories like *european rugby cup*. But as shown in subsequent columns of Table 2, after generalization top level *hockey* and *server* related categories are prominent. As $\beta$ increases, the rank of the categories *ice hockey* and *hockey* increases but the rank of *servers* decreases. This behavior is perhaps due to the presence of more *hockey* related terms such as 'playoff' and 'stats' in the document.

Table 2 does not show higher values of $\beta$ for space reasons. To summarize, as $\beta$ increases relevant concepts such as *ice hockey* and *automobiles* are pushed down in favor of broader concepts such as *sports* and *transportation*, and as $\beta$ approaches 1, overly general categories like *categories* and *human societies* take over.

D3 and D5 are included as examples of what can go wrong. In the case of D3, 'wordperfect' was not in the dictionary used to assign the cosine similarity scores. Therefore, one of the few significant words in the email was ignored. D5 shows the limits of our technique's robustness against failures in the inital topic assignment. Its top base concepts include *gospel music*, *gospel musicians*, and *grammy awards for gospel music*. These three categories form a small subgraph in the Wikipedia category graph (the first the immediate parent of the other two) that 'traps' the random walk, so that *gospel music* becomes the dominant category with $\beta = .5$. Obviously, the best solution is improving the base scores so that these concepts are given lower scores. We expect that a more unifrom Wikipedia category graph would also improve this sort of pathology, since more relevant categories would have higher base scores. On the other hand, there are steps we could take to make the algorithm more robust, and we are researching them.

### 3.2 Performance on an entire newsgroup

We want to show how the technique significantly homogenizes and generalizes the discovered concepts across an entire newsgroup. We coin the measurement '$top_N$' to show how common a topic is across an entire newsgroup. Let D be some data set. Let $\mathcal{T}$ be a *concept tagging* of of documents in D. That is, if $\mathcal{C}$ is category ontology, then $\mathcal{T}$ is function $\mathcal{T} : D \times \mathcal{C} \rightarrow [0, 1]$, and $\mathcal{T}(d_i, c_j)$ is an affinity of concept $c_j$ to document $d_i$. Note that this is a formalism of the first axiom of the base concept procedure we described in 1.3. We can define $top_N(d_i, \mathcal{T})$ as the N highest ranked concepts for a particular document according to the tagging $\mathcal{T}$. Finally, define $max_N$ of a collection to be the N most frequently occurring elements in the collection. From there, we can define the top n concepts for an entire data set:

$$top_N(D, \mathcal{T}) = max_N\{top_1(d_i, \mathcal{T}) : d_i \in D\}.$$

That is, we take the top-ranked concept for each document in D and put them all in a list and then find the N most frequent elements in the list. The dependence on $\mathcal{T}$ will be clear from context, so it is suppressed.

For selected news groups in 20 NG, Table 3 shows $top_5(D)$ along with the number of documents $d_i$ for which the concept was $top_1(d_i)$. That is, the table shows the number of documents associated with the top concepts of the selected news groups. It is apparent that many more documents have a common concept after generalization.

Continuing this line of reasoning, let's say that a set of concepts 'covers' a document if one of the concepts in the set is the top concept for the document. In Figure 1, we show the size of the smallest set of concepts that covers a certain percentage of selected news groups.

One could object to this evaluation measure on the following grounds: suppose the generalization is so aggressive that every document is tagged with a trivial concept. Specifically, in our case, it is possible for the generalization algorithm to tag a document with the

| | Base Concepts | | $\beta = .2$ | | $\beta = .5$ | |
|---|---|---|---|---|---|---|
| D1 | servers | .31 | ice hockey | .01 | ice hockey | .02 |
| | hockey at the summer olympics | .22 | servers | .01 | hockey | .01 |
| | defunct ice hockey leagues | .21 | ice hockey leagues | .01 | ice hockey leagues | .01 |
| | web server software | .21 | web server software | .01 | nhl players by team | .01 |
| | united states hockey hall of fame | .21 | microsoft server technology | .01 | field hockey | .01 |
| | hockey | .19 | field hockey | .01 | servers | .01 |
| | microsoft server technology | .18 | hockey at the summer olympics | .01 | nhl | .01 |
| | european rugby cup | .18 | nhl players by team | .01 | sports in canada | .01 |
| | ottawa senators (original) players | .17 | hockey | .01 | team sports | .01 |
| | hockey hall of fame | .17 | defunct ice hockey leagues | .01 | sports | .01 |
| D2 | new york city subway passenger equipment | .30 | automobiles | .01 | automobiles | .03 |
| | car classifications | .20 | car classifications | .01 | vehicles | .02 |
| | supercars | .19 | supercars | .01 | car classifications | .02 |
| | car rental | .16 | new york... | .01 | transportation | .02 |
| | mid-engined vehicles | .16 | auto racing | .01 | auto racing | .02 |
| | historic electric vehicles | .15 | bmw | .01 | road transport | .01 |
| | world war ii armored cars | .15 | bmw vehicles | .01 | vehicles by brand | .01 |
| | ferrari vehicles | .15 | luxury vehicles | .01 | luxury vehicles | .01 |
| | racing cars | .14 | mosler vehicles | .01 | automobile manufacturers | .01 |
| | open wheel racing | .14 | bugatti vehicles | .01 | supercars | .01 |
| D3 | sans-serif typefaces | .10 | metros in japan | .02 | email | .01 |
| | x window managers | .10 | osaka municipal subway stations | .01 | typefaces | .01 |
| | digital typography | .09 | sans-serif typefaces | .01 | metros in japan | .01 |
| | queensland prisons | .09 | x window managers | .01 | email clients | .01 |
| | pi | .09 | email clients | .01 | firearms | .01 |
| | microsoft windows | .09 | osaka municipal subway | .01 | x window system | .01 |
| | tra routes | .08 | tra routes | .01 | machine guns | .01 |
| | districts of bialystok | .08 | typefaces | .01 | railway stations in japan | .01 |
| | windowing systems | .08 | email | .01 | sans-serif typefaces | .01 |
| | moscow metro lines | .08 | pi | .01 | x window managers | .01 |
| D4 | muscular system | .06 | benzodiazepines | .02 | pharmacologic agents | .03 |
| | non-steroidal anti-inflammatory drugs | .06 | non-steroidal... | .02 | medicine | .02 |
| | muscle relaxants | .05 | pharmacologic agents | .02 | fluid mechanics | .02 |
| | non-newtonian fluids | .05 | fluid mechanics | .02 | human anatomy | .01 |
| | fluid mechanics | .04 | non-newtonian fluids | .02 | analgesics | .01 |
| | analgesics | .04 | fluid dynamics | .02 | medical specialties | .01 |
| | atc codes | .04 | muscular system | .02 | fluid dynamics | .01 |
| | anticoagulants | .04 | muscle relaxants | .01 | continuum mechanics | .01 |
| | over-the-counter substances | .04 | opioids | .01 | non-steroidal... | .01 |
| | opioids | .04 | analgesics | .01 | muscle relaxants | .01 |
| D5 | new testament apocrypha | .12 | new testament apocrypha | .03 | gospel music | .02 |
| | hiberno-saxon manuscripts | .12 | biblical scholars | .02 | new testament | .02 |
| | gospel music | .10 | gospel music | .02 | bible | .02 |
| | gospel musicians | .09 | biblical criticism | .02 | new testament apocrypha | .01 |
| | biblical criticism | .08 | hiberno-saxon manuscripts | .02 | christianity | .01 |
| | grammy awards for gospel music | .08 | new testament books | .01 | religious texts | .01 |
| | lost works | .08 | gospel musicians | .01 | biblical criticism | .01 |
| | biblical scholars | .08 | grammy awards for gospel music | .01 | christian texts | .01 |
| | gnosticism | .08 | new testament | .01 | new testament books | .01 |
| | new testament books | .07 | deuterocanonical books | .01 | biblical scholars | .01 |

Table 2: For the test documents given in Table 1 the above table shows how the top 10 generalized concepts vary with $\beta$.

root category, called "category." If every document were tagged with this concept, then by this measure, the algorithm would perform perfectly. One can look at Table 3 to see that this is not happening. The top concepts for each newsgroup are almost entirely distinct.

For rec.sport.baseball, the fifth top base concept was "Jewish media," and it was a top concept in 23 documents in the newsgroup. The top 5 newsgroup concepts cover only 22% of the articles. We con-

| | | | |
|---|---|---|---|
| rec.sports.baseball | Base | baseball pitching | 79 |
| | | baseball computer games | 71 |
| | | chicago cubs | 26 |
| | | major league designated hitters | 25 |
| | Gen | baseball | 366 |
| | | american league all-stars | 123 |
| | | computer and video games | 94 |
| | | national league all-stars | 62 |
| soc.religion.christian | Base | singular god | 68 |
| | | christian viewpoints | 65 |
| | | jesus | 56 |
| | | study bibles | 45 |
| | Gen | christianity | 307 |
| | | religious faiths, traditions, and movements | 88 |
| | | jesus | 54 |
| | | lgbt | 39 |
| rec.sports.hockey | Base | defunct ice hockey leagues | 47 |
| | | boston bruins coaches | 37 |
| | | power rangers | 35 |
| | | toronto maple leafs coaches | 34 |
| | Gen | ice hockey | 191 |
| | | computer and video games | 179 |
| | | sports | 77 |
| | | nhl | 59 |
| rec.autos | Base | new york city subway passenger equipment | 234 |
| | | oils | 32 |
| | | insurance companies of japan | 26 |
| | | automotive braking technologies | 26 |
| | Gen | automobiles | 433 |
| | | engines | 32 |
| | | oils | 27 |
| | | vehicles | 27 |
| sci.crypt | Base | clippers | 128 |
| | | computer keys | 82 |
| | | sound chips | 57 |
| | | block ciphers | 43 |
| | Gen | cryptography | 273 |
| | | clippers | 70 |
| | | algorithms | 58 |
| | | law | 34 |
| sci.space | Base | earth orbits | 66 |
| | | space shuttle program | 58 |
| | | space advocacy | 47 |
| | | gamma-ray telescopes | 25 |
| | Gen | space | 150 |
| | | space exploration | 143 |
| | | astronomy | 69 |
| | | hip objects | 16 |
| talk.politics.guns | Base | heavy machine guns | 81 |
| | | wildfires | 54 |
| | | firearm laws | 40 |
| | | bill clinton | 24 |
| | Gen | firearms | 181 |
| | | weapons | 64 |
| | | firefighting | 49 |
| | | bill clinton | 30 |
| talk.politics.mideast | Base | cities and towns in armenia | 111 |
| | | jews by country | 70 |
| | | political parties in palestine | 34 |
| | | israeli-palestinian conflict | 30 |
| | Gen | israel | 267 |
| | | judaic studies in academia | 61 |
| | | history of armenia | 51 |
| | | cyprus | 48 |

Table 3: Top 4 concepts (with the number of documents for which they are the top concept) for selected news groups before and after generalization ($\beta = .5$).

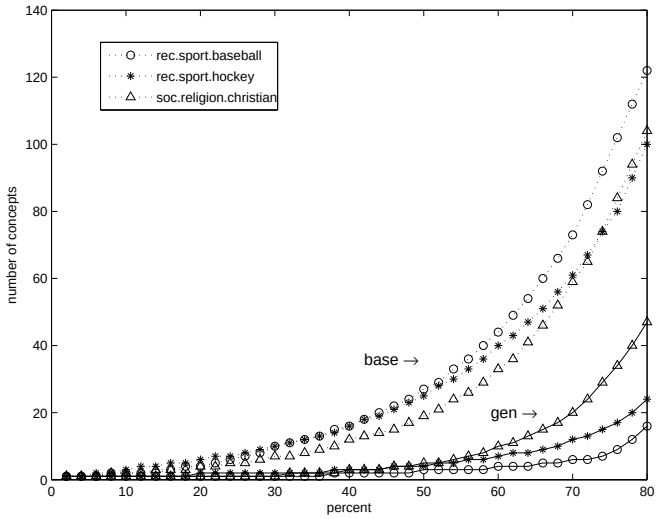


Figure 1: Smallest number of concepts required to cover the n$^{th}$ percentile of selected news groups before and after generalization ($\beta = .5$).

clude that the concept tagging by the base concept procedure included a lot of diversity, that is, a lot of topics were germane to only a few documents in the newsgroup, or were simply incorrect. After running the proposed algorithm (with $\beta = .5$), nearly all the concepts are germane to the newsgroup as a whole and over 70% of the documents are covered by the newsgroup's top 5. A similar analysis holds for rec.sport.hockey. Finally, the soc.religion.christian newsgroup provides an example where the base concepts are all germane. Nevertheless, the proposed algorithm is able to significantly homogenize the concepts across all the documents in this newsgroup as well.

From this evaluation, we conclude that the proposed technique is able to discover information about a particular document or a newsgroup as a whole that is not apparent from the base concepts tagging. One might dismiss this evaluation by saying the base concepts are so noisy that anything would be an improvement. Our response is twofold. First, we were able to achieve our results based on our base concepts procedure. Second, we would expect our results to be greatly improved with better base concept tagging.

# 4 Conclusion

We have proposed a new PageRank-based technique to discover generalized concepts from a document using an external category ontology such as Wikipedia. The articles associated with each category are used to perform the concept discovery. We also provide a mechanism to control how specific or how general the discovered concepts should be. In the absence of a standardized test data set, we evaluate our results using two different approaches — first, qualitative evaluation using selected documents and second, studying the homogeneity between the concepts identified for a particular document and the concepts identified for the entire set to which it belongs.

Our initial experiments with concept generalization shows that our approach shows promise in being able to extract the broad concept or topic associated with a document, even if the document is too short to contain many of the words common to that broad concept. In the process of carrying out our experiments, we identified several issues that would have to be addressed.

First, it is necessary to have a dictionary that contains all words that would be importance for identifying the concept of an article. Since many test documents in 20 NG are quite short, missing one significant word might be the difference between success and failure for that document. For example, 'Wordperfect' was a prominent word in example document D3, but it was not in our dictionary. The ideal dictionary, however, might be so large as to make computation intractable. An alternative is to consider byte grams (a collection of bytes). Moreover, [14] suggests that short byte grams will suffice.

Second, improving the initial base scores assigned to the document will improve the quality of the generalized concepts as well. First, many messages in 20 NG have spurious signature lines, and incremental improvement could be made by filtering these out. For instance, D3 has a signature line which includes "Naked Gun Part (Pi)." In Table 2 one can see that *pi* is a top base concept and *firearms* is one of the top generalized concepts. Generally, cosine scores are too dependent on the number of words present in the document and do not consider the co-occurrence of words. For example, for document $D5$ (see Table 2) a number of concepts related to gospel music were in the top base concepts, even though the co-occurrence of words such 'Essene' and 'syriac' suggests that the concept *biblical criticism* should have a higher score. One approach would be to use a classifier in the initial phase. But this step is complicated by the large number of category labels and documents. Moreover, a classifier ought to consider the hierarchical dependency between the classes, and, at first glance, building such a classifier is not easy.

Another issue arises from irregularities in the Wikipedia hierarchy. Some topics are split much more finely into categories than others, so these portions of the graph tend to trap the random walk. Even if most of the nodes have a low base score, the large number of nodes combined with the large number of interconnecting edges tends to overpower other, perhaps more germane, portions of the graph in the random walk. This issue was discussed earlier, when we considered D5 and the *gospel music* concept. By detecting such subgraphs, we could mitigate this problem.

# 5 Acknowledgments

This research was partially supported by NSF grant 0534286 and DARPA STTR grant W31P4Q-08-C-0242.

# References

[1] P. Avesani, F. Giunchiglia, and M. Yatskevich. A large scale taxonomy mapping evaluation. In *IWSC*, pages 67–81, Berlin, Germany, 2005. Springer-Verlag.

[2] P. Bouquet, L. Serafini, and S. Zanobini. Semantic coordination: a new approach and an application. In *ISWC '03*, Berlin, Germany, 2003. Springer-Verlag.

[3] A. Capocci, F. Rao, and G. Caldarelli. Taxonomy and clustering in collaborative systems: The case of the on-line encyclopedia wikipedia. *Europhysics Letters*, 81(2), 2008.

[4] K. Coursey and R. Mihalcea. Topic identification using wikipedia graph centrality. In *Proc. of Human*

*Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 117–120, Boulder, Colorado, June 2009. Association for Computational Linguistics.

[5] K. Coursey, R. Mihalcea, and W. Moen. Using encyclopedic knowledge for automatic topic identification. In *CoNLL-2009*, pages 210–218, Boulder, Colorado, June 2009. Association for Computational Linguistics.

[6] L. Denoyer and P. Gallinari. The Wikipedia XML Corpus. *SIGIR Forum*, 2006.

[7] E. Frank, G. W. Paynter, I. H. Witten, C. Gutwin, and C. G. Nevill-Manning. Domain-specific keyphrase extraction. In *IJCAI '99: Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, pages 668–673, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.

[8] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI '07*, 2007.

[9] E. Gabrilovich and S. Markovitch. Wikipedia-based semantic interpretation for natural language processing. *Journal of Artificial Intelligence Research*, 34:443–498, March 2009.

[10] S. Gollapudi and R. Panigrahy. Exploiting asymmetry in hierarchical topic extraction. In *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 475–482, New York, NY, USA, 2006. ACM.

[11] A. Hulth. Improved automatic keyword extraction given more linguistic knowledge. In *Proc. 2003 Conf. on Empirical methods in natural language processing*, pages 216–223, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

[12] Y. Labrou and T. Finin. Yahoo! as an ontology: using yahoo! categories to describe documents. In *CIKM '99*, pages 180–187, 1999.

[13] C. D. Manning, P. Raghaven, and H. Schütze. *Introduction to Information Retrieval*. Cambridge U. P., New York, 2008.

[14] J. E. Mason, M. Shepherd, and J. Duffy. Classifying web pages by genre: An n-gram approach. In *2009 IEEE/WIC/ACM Intl Conf on Web Intelligence and Intelligent Agent Technology*, volume 1, 2009.

[15] O. Medelyan and I. H. Witten. Thesaurus based automatic keyphrase indexing. In *JCDL '06: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, pages 296–297, New York, NY, USA, 2006. ACM.

[16] O. Medelyan, I. H. Witten, and D. Milne. Topic indexing with wikipedia. In *AAAI WikiAI workshop*, 2008.

[17] R. Mihalcea and A. Csomai. Wikify!: linking documents to encyclopedic knowledge. In *CIKM '07*, pages 233–242, New York, NY, USA, 2007. ACM.

[18] V. Nastase and M. Strube. Decoding Wikipedia categories for knowledge acquisition. In *AAAI*, pages 1219–1224, 2008.

[19] P. Schonhofen. Identifying document topics using the Wikipedia category network. In *WI '06*, pages 456–462, Washington, DC, USA, 2006. IEEE Computer Society.

[20] M. Strube and S. P. Ponzetto. Wikirelate! computing semantic relatedness using Wikipedia. In *AAAI '06*, pages 1419–1424, 2006.

[21] Z. S. Syed, T. Finin, and A. Joshi. Wikipedia as an ontology for describing documents. In *ICWSM '08*, 2008.

[22] S. Tiun, R. Abdullah, and T. E. Kong. Automatic topic identification using ontology hierarchy. In *CICLing '01*, pages 444–453, London, UK, 2001. Springer-Verlag.

[23] P. D. Turney. Learning algorithms for keyphrase extraction. *Inf. Retr.*, 2(4):303–336, 2000.

[24] C. yew Lin. Knowledge-based automatic topic identification. In *Proc. of The 33rd Annual Meeting of the Association for Computational Linguistics '95*, pages 308–310, 1995.