Wikipedia as an Ontology for Describing Documents

Zareen Saba Syed, Tim Finin and Anupam Joshi
University of Maryland, Baltimore County, Baltimore MD 21250
zarsyed1@umbc.edu, finin@cs.umbc.edu, joshi@cs.umbc.edu

Abstract

Identifying topics and concepts associated with a set of documents is a task common to many applications. It can help in the annotation and categorization of documents and be used to model a person's current interests for improving search results, business intelligence or selecting appropriate advertisements. One approach is to associate a document with a set of topics selected from a fixed ontology or vocabulary of terms. We have investigated using Wikipedia's articles and associated pages as a topic ontology for this purpose. The benefits are that the ontology terms are developed through a social process, maintained and kept current by the Wikipedia community, represent a consensus view, and have meaning that can be understood simply by reading the associated Wikipedia page. We use Wikipedia articles and the category and article link graphs to predict concepts common to a set of documents. We describe several algorithms to aggregate and refine results, including the use of spreading activation to select the most appropriate terms. While the Wikipedia category graph can be used to predict generalized concepts, the article links graph helps by predicting more specific concepts and concepts not in the category hierarchy. Our experiments demonstrate the feasibility of extending the category system with new concepts identified as a union of pages from the page link graph.

Introduction

Characterizing what a document is "about" is a task common to many applications, including classification, retrieval, modeling a user's interests, and selecting appropriate advertisements. The work we report on in this paper was motivated by the requirements of the following application, which is under development.

> A team of analysts is working on a set of common tasks, with each analyst focused on several different areas and working sometimes independently and sometimes in a tightly coordinated group. Collaboration in such a setting is enhanced if the individual analysts maintain an awareness of what their colleagues have been working on. As new members join the team or return to it after a temporary assignment or vacation, it is important for them to acquire the context of who has been working on or is interested in what. A way to describe the topics on which an analyst focuses is through an analysis of the documents, or portions of documents that she has been reviewing, reading or writing. In short, if we know what she has been reading, we have a good handle on what she is working on.

One general approach to describing what a document is about is to use statistical techniques to describe the words and phrases it contains. This is the basis of information retrieval, which has had enormous practical success. Another approach is to tag the document with relevant terms that represent semantic concepts important to the document. This is typically used in information science using terms from a standard classification hierarchy such as the Dewey Decimal System (Dewey 1990) or ACM Computing Classification System (Coulter et al. 1998). More recently, many Web 2.0 systems have allowed users to tag documents and Web resources with terms without requiring them to come from a fixed vocabulary. In a social media context (e.g., del.icio.us or Flickr) an implicit ontology of tags can emerge from the community of users and subsequently influence the tags chosen by individuals, reinforcing a notion of a common ontology developed by the community.

An advantage of using the "ontology" approach, whether based on a designed or emergent ontology, is that the terms can be explicitly linked or mapped to semantic concepts in other ontologies and are thus available for reasoning in more sophisticated language understanding systems such as OntoSem (Nirenburg et al. 2004) and Powerset, or specialized knowledge-based systems, or in Semantic Web applications.

Using the traditional approach of a controlled, designed ontology has many disadvantages beginning with the often difficult task of designing and implementing the ontology. Once that it done, it must be maintained and modified, an important process in domains where the underlying concepts are evolving rapidly. ACM's CCS, for example, undergoes periodic reorganization and redesign and yet as a classification of computer science concepts, it always seems to be out of date or even quaint. As a final problem, consider the process a person must follow in assigning ontology terms to a document. She has to be familiar with all of the possible choices or have some way to browse or search through them. She has to understand what each of the terms means, either the original meaning intended by the ontology designer or the possibly different current meaning as used by her community. Finally, she has to select the best set of terms from among the many relevant choices the ontology may present to her.

The use of an implicit ontology emerging from the tagging choices of a community of individuals solves some of these problems, but also has significant disadvantages. Some of these are inherent and others are being addressed in the research community and may ultimately admit good

solutions. These problems are worth addressing because the result will be an ontology that (1) represents a consensus view of a community of users and (2) is constructed and maintained by the community without cost to any organization. It remains unclear how the terms in such an ontology should be organized structurally, understood informally by end users, or mapped to a more formal ontology such as Cyc (Lenat 1995) or popular Semantic Web ontologies like FOAF (Ding et al. 2005).

We are developing a system that is a blend of the two approaches based on the idea of using Wikipedia as an ontology. Specifically, each non-administrative Wikipedia page is used as a term in an ontology. These include Wikipedia articles describing individuals (Alan Turing), concepts (Emissions trading), locations (Barbados), events (collapse of the World trade Center), and categories (microbiology). Using Wikipedia as an ontology has many advantages: it is broad and fairly comprehensive, of generally high quality, constructed and maintained by tens of thousands of users, evolves and adapts rapidly as events and knowledge change, and free and "open sourced". Moreover, the meaning of any term in the ontology is easy for a person to understand from the content on the Web page. Finally, the Wikipedia pages are already linked to many existing formal ontologies though efforts like DBpedia (Auer et al. 2007) and Semantic MediaWiki (Krotzsch et al. 2006) and in commercial systems like Freebase and Powerset.

The underlying concept of an article cannot be assessed by merely considering the words that appear in that article, in addition to that, finding out if two articles are conceptually related is an even more challenging problem and requires a lot of background domain knowledge, common sense as well as information about the context. Humans have the inborn ability to relate concepts semantically however it is still a very difficult problem for computers, which can be made easier by augmenting background domain knowledge for such tasks, which would certainly improve the accuracy and quality of prediction. Wikipedia proves to be an invaluable source for such background domain knowledge.

Wikipedia

Wikipedia is a freely available online encyclopedia developed by a community of users. Wikipedia is growing exponentially and new content is being added to it daily by users around the globe. This encyclopedia comprises of millions of articles. The corpus is composed of several collections in different languages such as: English, French, German, Dutch, Chinese, Spanish, Arabic and Japanese. Each collection is a set of XML documents built using Wikipedia.

Documents of the Wikipedia XML collections are organized in a hierarchy of categories defined by the authors of the articles. The Wikipedia category and article network has been studied in detail with respect to different graph properties. The Wikipedia category system is a taxonomy for arranging articles into categories and sub-categories. However, this taxonomy is not a strict hierarchy or tree of categories, but allows multiple categorizations of topics simultaneously, i.e., some categories might have more than one super-category. It is shown that Wikipedia's category system is a thesaurus that is collaboratively developed and used for indexing Wikipedia articles (Voss 2006). The articles with in Wikipedia are inter-linked. However, these links do not impose any sub-category or super-category relationships. It has been observed that the Wikipedia article links graph generally resembles the World Wide Web graph (Zlatic et al. 2006).

Spreading Activation

Spreading Activation is a technique that has been widely adopted for associative retrieval (Crestani 1997). In associative retrieval the idea is that it is possible to retrieve relevant documents if they are associated with other documents that have been considered relevant by the user. In Wikipedia the links between articles show association between concepts of articles and hence can be used as such for finding related concepts to a given concept. The algorithm starts with a set of activated nodes and in each iteration the activation of nodes is spread to associated nodes. The spread of activation may be directed by addition of different constraints like distance constraints, fan out constraints, path constraints, threshold etc. These parameters are mostly domain specific.

In this study we consider a Wikipedia category as a representative of a generalized concept. The title of a Wikipedia article may be considered as a specific or specialized concept. The links between different articles are considered as links between different concepts. We have implemented different heuristics to use the Wikipedia article texts, category network and page links graph for predicting concepts related to documents.

This paper is organized as follows: A brief review of literature related to the use of Wikipedia for information retrieval is discussed in section II. In section III we discuss our implementation details as well as our parameters for network spreading algorithm that is used for associative information retrieval. Section IV described the results of some preliminary experiments. In section V we present results of an evaluation our method and in section VI we discuss the results of our experiments. Section VII concludes the paper and briefly describes suggests directions for future work.

Related Work

The depth and coverage of Wikipedia has attracted the attention of researchers who have used it as a knowledge resource for several tasks, including text categorization (Gabrilovich et al. 2006), co-reference resolution (Strube et al. 2006), predicting document topics (Schonhofen 2006), automatic word sense disambiguation (Mihalcea 2007), searching synonyms (Krizhanovsky 2006) and computing semantic relatedness (Strube et al. 2006, Gabrilovich et al. 2007, Milne 2007). To the best of our

knowledge, Wikipedia has not yet been directly used to predict concepts that characterize a set of documents.

While this is similar to the task of assigning documents to a class or category, it differs in a significant way. In categorization, the task is to predict the category of a given document however, predicting common concepts for a set of documents may include documents belonging to very different categories but having some concept in common. For example a user searching for information related to growing a flowering plant may consider reading different articles on seeds, fertilizers, herbicides, manure, gardening etc, all these articles may belong to different categories yet share a common concept that all are related to the plant. However, in certain cases in which the set of documents belong to the same category, we may be able to introduce the predicted common concept as a new sub-category.

We find our problem very similar in direction to computing semantic relatedness between concepts with the addition that we focus on predicting a concept that is common as well as semantically related to a set of documents. In this section we give a brief review of related work.

The initial work done on employing Wikipedia for computing semantic relatedness was by Strube and Ponzetto and realized in a system named WikiRelate! (Strube et al. 2006). They used information from Wordnet, Wikipedia and Google in computing degrees of semantic similarity and reported that Wikipedia outperforms Wordnet. However, they obtained the best results when evidence from all three resources was integrated. They used different measures for computing semantic relatedness including measures based on paths, information content, and text overlap.

Gabrilovich and Markovich used concept space derived from Wikipedia to compute semantic relatedness between fragments of natural language text, and reported the performance to be significantly better than other state of the art methods (Gabrilovich et al. 2007). They named their approach "Explicit Semantic Analysis" (ESA) as they use concepts that are explicitly defined by users. Their method employs machine learning technique to represent the meaning of a text fragment as a weighted vector of concepts derived from Wikipedia. Relatedness is then measured through the comparison of concept vectors using conventional metrics such as cosine similarity.

The success of their experiments gives support to our method, which also initially utilizes the Wikipedia concept space, although in a different manner. Instead of using machine learning techniques, we directly compute the related concepts based on the cosine similarity between the input document and Wikipedia articles and then use those concepts as our initial activated nodes in spreading activation. The key difference is that we are not interested in merely finding the semantic relatedness between documents but in finding a semantically related concept that is also common to a set of documents.

Wikipedia Link Vector Model is an approach that is similar to ESA that eliminates the need for processing Wikipedia article text (Milne 2007). This method computes the semantic relatedness between terms based on the links found in their corresponding Wikipedia articles. The reported results, however, give less accuracy than ESA.

Methodology

We downloaded the Wikipedia XML snapshot of 4 November 2006 and extracted 2,557,939 Wikipedia articles. The text of each article was indexed using the Lucene text search engine library (Gospodnetic et al. 2004) under the standard configuration. We ignored the history, talk pages, user pages, etc. We also downloaded the Wikipedia database tables in order to create the category links graph and the article links graph. Major administrative categories (e.g., "All articles with unsourced statements") were identified and removed from the category links graph. Any remaining administrative categories appearing in the prediction results were excluded. We implemented three different methods for our study, which are described and discussed below.

Method 1: Article Text

In the first method we use the test document or set of related documents as search query to the Lucene Wikipedia index. After getting top N matching Wikipedia articles (based on cosine similarity) for each document in the set, we extract their Wikipedia categories and score them based on two simple scoring schemes.

- In scoring scheme one we simply count the number of times that each Wikipedia category was associated with one of the N results.
- In scoring scheme two we take into account the cosine similarity score between the test document and matching Wikipedia articles. The score for each category is the sum of the cosine similarity scores of the matching articles that are linked to the category.

Method 2: Text and Categories with Spreading Activation

In the second method we also use the Wikipedia category links network for prediction of related concepts. We take the top N Wikipedia categories predicted as a result of method one scoring scheme one and use them as the initial set of activated nodes in the category links graph. After 'k' pulses of spreading activation, the category nodes are ranked based on their activation score.

Activation Function:

Node Input Function: $I_j = \sum_i O_i$

Node Output Function: $O_j = \frac{A_j}{D_j * k}$

Where the variables are defined as:

O_i : Output of Node i connected to node j

A_j : Activation of Node j

k : Pulse No.
D_j : Out Degree of Node j

Method 3: Text and Links with Spreading Activation

In the third method we take the top N matching Wikipedia articles (based on cosine similarity) to each test document as the initial set of activated nodes in the article links graph. During our preliminary experiments we observed that there were many irrelevant links between articles based on the fact that a title of one article appears as a word in the other for example, an article that mentions the name of a country (e.g., Canada) often has a link to the article on that country even though the country is mentioned in passing and is unrelated to the article's topic.

Hence to refine the links in the article links graph we filter out all links between documents whose cosine similarity scores are below a threshold (e.g., 0.4) so that the spreading activation would be more directed. We use three kinds of node activation functions for spreading activation. The objective of using three different activation functions was to see if there is any difference in the prediction results through each function.

Activation Function 1:

Node Input Function: $I_j = \sum_i O_i w_{ij}$

Node Output Function: $O_j = \frac{A_j}{k}$

Activation Function 2:

Node Input Function: $I_j = \sum_i O_i$

Node Output Function: $O_j = 1$

Activation Function 3:

Node Input Function: $I_j = \sum_i O_i$

Node Output Function: $O_j = \frac{A_j}{k}$

Where the variables are defined as:

O_i : Output of Node i connected to node j
A_j : Activation of Node j
w_{ij}: Weight on edge from node i to node j
k : Pulse No.
D_j : Out Degree of Node j

Experiments and Results

We conducted three different kinds of experiments. Our first experiment was focused at simply observing how well the Wikipedia categories represent concepts in individual test documents. For this we ran experiments using methods one and two. The second experiment was an extension to the first experiment in that it included a set of test documents rather than individual documents. The third experiment was targeted towards finding if a common concept could be predicted for a set of documents using the article links graph given that the concept is not already defined as a Wikipedia category.

Ex 1: Predicting the topic of a single document using Methods 1 and 2

In this experiment we took several articles representing various subjects and areas directly from Internet and used methods 1 and 2 to predict concepts related to individual articles. The results of the top three predictions in order of their rank for a few of those articles are given in table 1. We also give the actual titles of those test articles to evaluate the predictions.

The top most ranked prediction using both methods and scoring schemes in most of the cases match the title of the document or concept related to the title of the test document. We observed that the results don't significantly differ using the different methods and scoring schemes, however using spreading activation either results in the same prediction or a prediction of a more generalized concept that is evident in the results. In case of document 3, using three pulses for spreading activation resulted in prediction of a very generalized category named "Main topic classifications". Since the category graph is directed, i.e., from sub-categories to super-categories, it is expected that increasing the number of pulses will result in spreading activation to more generalized categories.

Ex 2: Predicting the topic of a set of documents using Methods 1 and 2

In this experiment, two test cases were run. For the first test case we took a set of ten documents related to Genetics from the Internet and tried to predict a common or general concept covering all documents. For each document in the test set, the ten top matching Wikipedia articles were retrieved resulting in initial activation of 100 nodes for spreading activation in category links graph.

The results of this experiment, shown in Table 3, are also very encouraging and a related concept common to all is predicted in almost all cases. We observe that increasing the spreading activation pulses results in prediction of more generalized categories. For example, if we consider the top most ranked predictions, in case of method 1 the prediction is "Genetics" however, in case of spreading activation with 2 pulses the prediction is "Biology" and with three pulses the prediction is "Nature" which is an even broader concept than biology.

Table 1: Concept prediction results for a single test document using Method 1 and Method 2

Sr No.	Test Document Title	Method 1 Scoring Scheme 1	Method 1 Scoring Scheme 2	Method 2 Pulses=2	Method 2 Pulses=3
1	Geology	"Geology" "Stratigraphy" "Geology_of_the_United_Kingdom"	"Geology" "Stratigraphy" "Science_occupations"	"Earth_sciences" "Geology" "Physical_sciences"	"Earth_sciences" "Natural_sciences" "Physical_sciences"
2	Atomic Bombings of Nagasaki	"Atomic_bombings_of_Hiroshima_and_Nagasaki" "Manhattan_Project" "Nuclear_warfare"	"Atomic_bombings_of_Hiroshima_and_Nagasaki" "Manhattan_Project" "Nuclear_warfare"	"Atomic_bombings_of_Hiroshima_and_Nagasaki" "Warfare_by_type" "Manhattan_Project"	"Wars_by_country" "Military_history_of_the_United_States" "Nuclear_technology"
3	Weather Prediction of thunder storms (taken from CNN Weather Prediction)	"Weather_Hazards" "Winds" "Severe_weather_and_convection"	"Weather_Hazards" "Current_events" "Types_of_cyclone"	"Meterology" "Nature" "Weather"	"Main_topic_classifications" "Fundamental" "Atmospheric_sciences"

Table 2: Titles of documents in the test sets

Test Set 1	Test Set 2	Test Set 3
1. Basic Genetics 2. Exceptions to Simple Inheritance 3. Mendel's Genetics 4. Probability of Inheritance 5. Basics of population genetics 6. Chromosomes 7. Coat Color Genetics 8. Genes 9. Inbreeding and Linebreeding 10.Structure of DNA	1. azithromycin 2. cephalexin 3. ciprofloxacin 4. clarithromycin 5. doxycycline 6. erythromycin 7. levofloxacin 8. ofloxacin 9. tetracycline 10. trimethoprim	1. Crop_rotation 2. Permaculture 3. Beneficial_insects 4. Neem 5. Lady_Bird 6. Principles_of_Organic_Agriculture 7. Rhizobia 8. Biointensive 9.Intercropping 10. Green_manure

Table 3: Common concept prediction results for a set of documents

Test Set	Method 1, Scoring Scheme 1	Method 1, Scoring Scheme 2	Method 2, Pulses 2	Method 2, Pulses 3
1.	"Genetics" "Classical_genetics" "Population_genetics"	"Genetics" "Classical_genetics" "Population_genetics"	"Biology" "Genetics" "Life"	"Nature" "Academic_disciplines" "Main_topic_classification"
2.	"Antibiotics" "Macrolide_antibiotics" "Organofluorides"	"Macrolide_antibiotics" "Antibiotics" "Organofluorides"	"Medicine" "Antibiotics" "Medical_specialties"	"Biology" "Human" "Health_sciences"
3.	"Agriculture" "Sustainable_technologies" "Crops"	"Agriculture" "Sustainable_technologies" "Crops"	"Skills" "Applied_sciences" "Land_management"	"Knowledge" "Learning" "Industries"

Table 4: Common concept prediction results for a set of documents related to "Organic farming" using Method 3 with different pulses and activation functions.

Pulses	Activation Function 1	Activation Function 2	Activation Function 3
1	"Organic_farming" "Sustainable_agriculture" "Organic_gardening"	"Organic_farming" "Sustainable_agriculture" "Agriculture"	"Permaculture" "Crop_rotation" "Green_manure"
2	"Organic_farming" "Permaculture" "Crop_rotation"	"Permaculture" "Organic_farming" "Sustainable_agriculture"	"Organic_farming" "Sustainable_agriculture" "Organic_gardening"

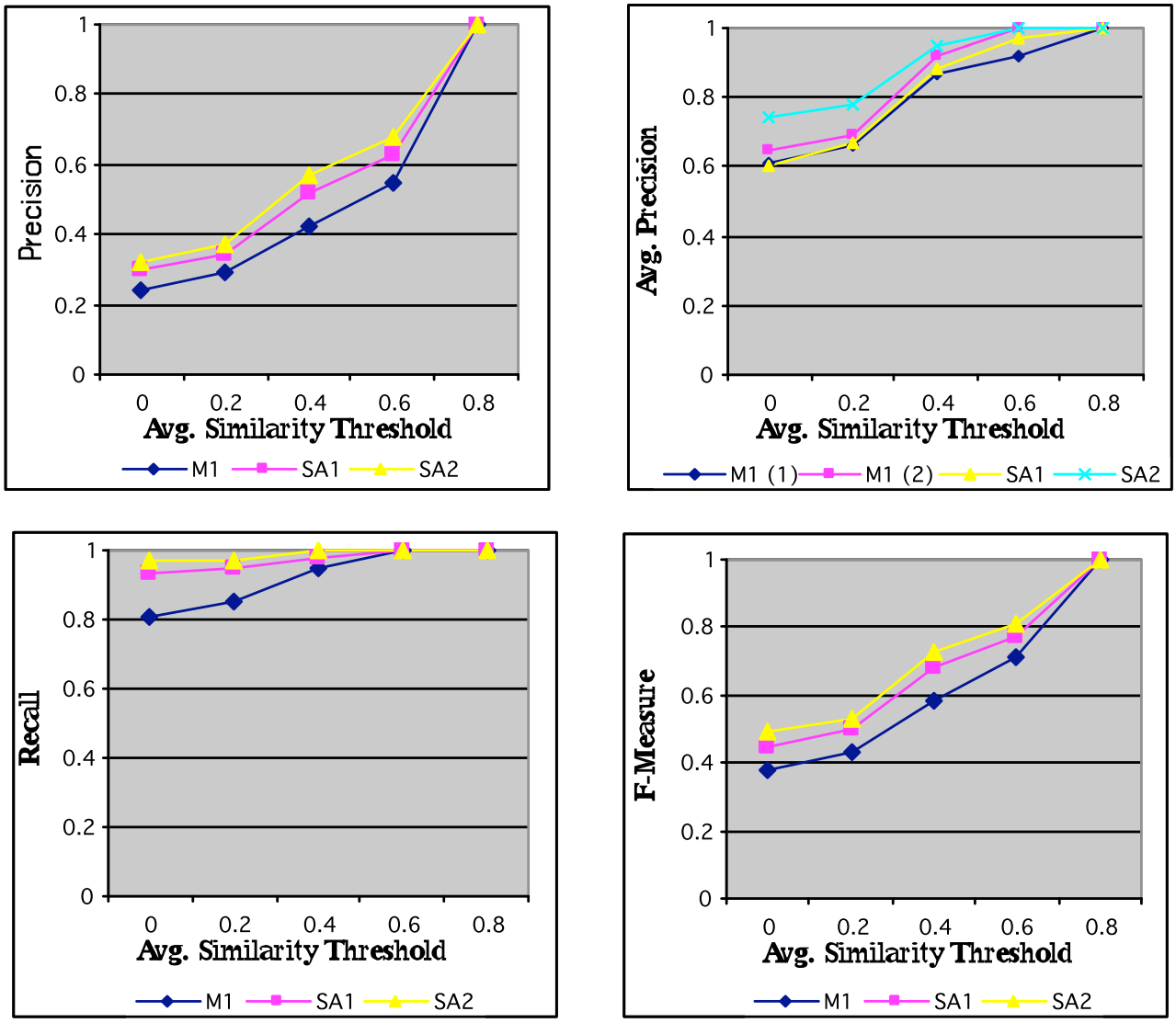

Figure 1. These graphs show the precision, average precision, recall and f-measure metrics as the average similarity threshold varies from 0.1 to 0.8. Legend label M1 is method 1, M1(1) and M(2) are method 1 with scoring schemes 1 and 2, respectively, and SA1 and S2 represent the use of spreading activation with one and two pulses, respectively.

This same experiment was repeated for the second test case, for which we took ten articles related to different antibiotics from internet. The top most ranked predictions using method 1 are "Antibiotics" using scoring scheme 1 and "Macrolide antibiotics" using scoring scheme 2. In case of Method 2 with spreading activation the top ranked predictions are "Medicine" with two pulses and "Biology" with three pulses. It is again observed that increasing the pulses results in prediction of a more generalized concept.

Ex 3: Predicting Common Concepts using Page Links and Categories

The objective of this experiment was to see if it is possible to predict a concept common to a set of Wikipedia articles themselves, given that the concept is not already represented as a Wikipedia category by using the article text and links. For this experiment we picked a set of related Wikipedia articles belonging to different categories but sharing some common concept as input. We picked different Wikipedia articles related to the concept of "Organic Farming" which is not represented as a category in Wikipedia. We used all three proposed methods to see if method 3, which also uses the page links information, can predict a common concept that is more specialized than the concepts predicted using methods 1 and 2.

The top ranked predictions using method 1 and 2 (Table 3) are "Agriculture", "Skills" and "Knowledge" whereas by using method 3 and different activation functions (Table 4), the top ranked predictions are "Organic farming" and "Permaculture" (Permaculture: means Permanent

Agriculture, which is also a related concept to Organic farming). These results show that it is possible to predict concepts common to a set of documents belonging to different categories by utilizing the article link structure of Wikipedia. We further observe that using Method 1 and 2 the best common concept that is predicted is very generalized i.e., "Agriculture" whereas by utilizing the article links we are able to predict a more specialized common concept. Using method 3 we can analyze Wikipedia as a whole to introduce new sub-categories with in Wikipedia and aid in enriching its category hierarchy. We used three activation functions in our method 3 however; we do not note any significant difference in predictions using the different activation methods for this experiment.

Empirical Evaluation

The experiments described in the previous section produced results that were encouraging, but serve only as an informal evaluation. The scale was small and the accuracy of the results were based on our own, possibly biased, judgments. We designed and ran a more formal evaluation by creating a test set of 100 articles randomly from Wikipedia. We removed references to those articles from our Wikipedia index, article links graph and category graph. We then used our system to find related articles and categories for each of the 100 articles. The results were compared to the actual categories and article links found in Wikipedia, which we took to be the "ground truth", wielding measures of precision, recall and F-measure.

For evaluating the category prediction, for each Wikipedia test article we retrieved top ten similar articles from Wikipedia index based on cosine similarity between the documents. We took the average of the cosine similarity score between the test article and the top ten similar Wikipedia articles and sorted the test articles based on that score. We computed precision, average precision, recall and F-measure at different similarity score thresholds for all methods. For example, at 0.5 average similarity threshold we computed all metrics for the subset of test documents that had a score of greater than or equal to 0.5. For computing these metrics we included the top three level categories to the actual categories of the test documents so that if our method predicts a category that is a super-category at a distance of three then we consider it to be an accurate prediction.

Figure 1 shows our results. We observed that higher the average similarity scores the better the precision, average precision and recall for all methods. A comparison of the different methods using the F-measure metric shows that the method using spreading activation with two pulses (SA2) almost always performs better than other methods at different average similarity thresholds and also for the test document set as a whole. Measuring the average precision gives us an idea of our ranking schemes. We observed that in all cases the average precision is better than the precision for all methods indicating that our scoring scheme gives a higher score to the relevant results. The best average precision is given by method SA2 and is always higher

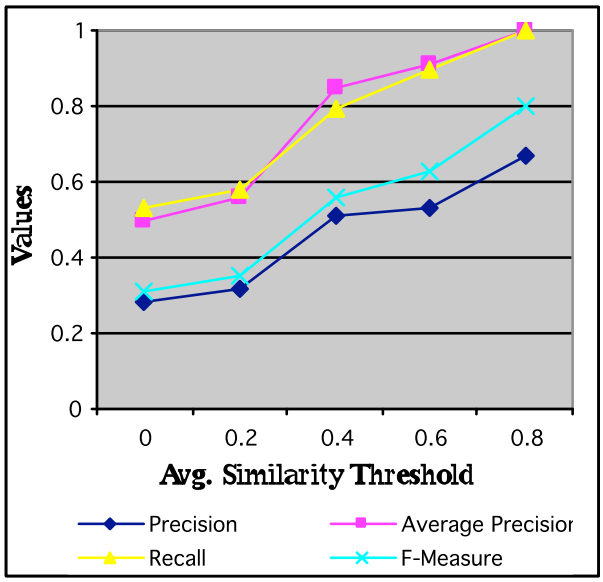

Figure 2. In the category prediction task, precision, average precision and recall improve at higher similarity thresholds, with average precision remaining higher than precision, indicating that our ranking scheme ranks relevant links higher than irrelevant links.

than other methods. Incase of Recall, SA2 gives highest recall at all thresholds. We also observe that M1(2) gives higher average precision than M1(1) hence, showing that scoring scheme 2 based on cosine similarity is superior to scoring scheme 1 based on number of occurrences. M1(1) also outperforms SA1 incase of average precision however, it is always lower than for SA2.

To evaluate our method for related concept prediction using Wikipedia article text and links, we used our test articles and removed their references from Wikipedia page links graph. Using our method we predicted the links of those articles and compared them with the actual links within Wikipedia using precision, average precision and recall. For each test article we retrieved top five similar articles and ran spreading activation with one pulse and activation function number 2 with unit edge weights. Figure 2 shows our results for related concept prediction. We again observed that similar to the category prediction case the precision, average precision and recall improve at higher average similarity thresholds. However, at lower similarity thresholds the precision and recall are greatly affected. We also observe that the average precision is always significantly higher than precision, indicating that our ranking scheme ranks relevant links higher than irrelevant links.

Discussion

We have conducted three sets of experiments and also evaluated our methods using Wikipedia articles themselves. In the first set of experiments we only utilized the Wikipedia page texts to predict the category or concept related to a document. We gave each test document as in-

put to Wikipedia articles index and got ten similar Wikipedia articles. We utilized the category information related to the matching articles to predict the category or concept of the test document using different scoring schemes. We experimented with a few documents and observed that the prediction was satisfactory for all of them. We also repeated the experiments with a group of documents related to a particular concept or topic instead of a single document and found the results to be encouraging in predicting the category of a group of related documents.

In the second set of experiments, in addition to using the Wikipedia article text we also applied spreading activation algorithm on the category links graph. The purpose of applying spreading activation was to find out if we could extract a generalization of the concept or a common concept presented in the test document or set of test documents. We observed that depending on the input parameters of spreading activation, it helped in predicting nodes representing a broader or more generalized concept as compared to the initial prediction of concept. This method was observed to be useful in predicting the super-categories or super-concepts of the test documents.

In the third set of experiments we also included the article links information. The purpose of the experiment was to investigate if it is possible to predict a common concept for a set of documents given that the concept is not already represented as a Wikipedia category. Our general observation was that the concepts that are sufficiently represented in Wikipedia usually have a category associated with them, however, there may be certain cases where several pages may be related to a particular concept and that concept may not be represented as a category. To study this we took few such examples from Wikipedia and ran spreading activation on the article links graph to predict a related concept to a set of documents.

The results of experiments for predicting more specialized concepts related to a group of documents were also encouraging. Such a concept could be considered as representing a specialized topic related to a set of documents in contrast to a generalized topic or category. If the group of documents under consideration belongs to the same category then the predicted specialized concept could be used in defining a new Wikipedia sub-category whereas, if the group of documents does not belong to the same category then the specialized concept could be used in defining a new relationship between those documents. For example, if we have an article related to a person and another article related to a location, we might be able to predict that the particular person and location are related to each other given a particular event which involved that person and occurred at the respective location however, we would not want to classify that person and location under that event.

An interesting application of the different methods that we have implemented and evaluated is that these methods could be used in recommending the categories and article links for new Wikipedia articles, or even in automatically building an enterprise Wiki from a given corpus by running our algorithms that utilize the category and article links information already present in Wikipedia.

Since we are currently employing Wikipedia as our knowledge base, predicting common concept to a set of documents is highly dependent on different factors inherent to Wikipedia:

- To what extent is the concept represented in Wikipedia: For example, there exists a category for the fruit "apple" however there is no category for "mango" since apple and its different varieties are discussed in detail in Wikipedia whereas for mango such information is limited to a few varieties.
- Presence of links between semantically related concepts: Since Wikipedia is developed by its users and not necessarily by experts hence the author of an article may not be able to add links to all other semantically related articles, and also doing that manually is infeasible in itself.
- Presence of links between irrelevant articles: Articles may be linked to other Wikipedia articles irrelevant to their topics or concepts. For example articles mentioning a name of a country may be linked to that country's Wikipedia page. An article that mentions a term may be liked to the article defining and giving details on that term.

Hence the accuracy of our method is largely dependent on the above three factors. However, we have shown through our evaluation that the greater the similarity between a test article and its similar Wikipedia articles the better the prediction. Therefore the average similarity score may be used to judge the accuracy of prediction. For factors 2 and 3 related to the presence and absence of semantically related links between articles we could use the existing semantic relatedness measures to introduce additional links between semantically related articles or to filter out links between irrelevant articles.

Conclusion

In this paper we described the use of Wikipedia and spreading activation to find generalized or common concepts related to a set of documents using the Wikipedia article text and hyperlinks. We started our experiments with the prediction of concepts related to individual documents, extended them to predict concepts common to a set of related documents, and used the text and links of uncategorized Wikipedia articles to predict extant Wikipedia articles to serve as a category term. We have discussed the results of our experiments and have also evaluated them using random articles from Wikipedia itself.

Our experiments show that it is possible to predict concepts common to a set of documents by using the Wikipedia article text and links. We have also discussed some possible solutions for improving our results. Where earlier work has been directed towards computing semantic relatedness between text fragments, we have focused on a more challenging task of finding semantically related concepts common to a set of documents. We are also currently working on applying machine learning techniques to clas-

sify links between Wikipedia articles. The results can be used as another source of evidence to predict the semantic "type" of an article (e.g., person, event, location) and to control the flow of spreading activation semantically.

References

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. 2007. DBpedia: A Nucleus for a Web of Open Data. Proceedings of the Sixth Int'l Semantic Web Conference, Springer, November 2007.

Coulter, N. et al. 1998. Computing Classification System 1998: Current Status and Future Maintenance Report of the CCS Update Committee. Computing Reviews, 1998, ACM Press New York, NY, USA.

Crestani, F. 1997. Application of Spreading Activation Techniques in Information Retrieval. Artificial Intelligence Review, 1997, v. 11; n. 6, 453-482.

Dewey, M. 1990. Abridged Dewey Decimal Classification and Relative Index, Forest Press.

Ding, L., Zhou, L., Finin, T., and Joshi, A. 2005. How the Semantic Web is Being Used: An Analysis of FOAF Documents. Proceedings of the 38th Annual Hawaii International Conference on System Sciences.

Gabrilovich, E., and Markovitch, S. 2006. Overcoming the brittleness bottleneck using Wikipedia: Enhancing text categorization with encyclopedic knowledge. Proceedings of the Twenty-First National Conference on Artificial Intelligence. AAAI'06. Boston, MA.

Gabrilovich, E., and Markovitch, S. 2007. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis, Proc. of the 20th International Joint Conference on Artificial Intelligence (IJCAI'07), 6-12.

Gospodnetic, O., and Hatcher,E. 2004. Lucene in Action, Manning Publications, December, 2004.

Krizhanovsky, A. 2006. Synonym search in Wikipedia: Synarcher. URL:http://arxiv.org/abs/cs/0606097v1

Krotzsch, M., Vrandecic, D. and Volkel, M. 2006. Semantic MediaWiki. Proceedings of the Fifth International Semantic Web Conference, pp 935-942, Springer, November 2006.

Lenat, D. B. 1995. CYC: a large-scale investment in knowledge infrastructure. Communications of the ACM, v38, n11, pp. 33-38, 1995, ACM Press New York, NY, USA.

Mihalcea, R. 2007. Using Wikipedia for Automatic Word Sense Disambiguation. Proc NAACL HLT. 196-203.

Milne, D. 2007. Computing Semantic Relatedness using Wikipedia Link Structure. Proceedings of the New Zealand Computer Science Research Student conference (NZCSRSC'07), Hamilton, New Zealand.

Nirenburg, S., Beale, S., and McShane, M. 2004. Evaluating the Performance of the OntoSem Semantic Analyzer. Proceedings of the ACL Workshop on Text Meaning Representation.

Powerset: http://www.powerset.com/

Schonhofen, P. 2006. Identifying Document Topics Using the Wikipedia Category Network. Proc. 2006 IEEE/WIC-/ACM International Conference on Web Intelligence. 456-462, 2006. IEEE Computer Society, Washington, DC, USA.

Strube, M., and Ponzetto, S.P. 2006. Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (2006). Association for Computational Linguistics Morristown, NJ, USA.

Strube, M., and Ponzetto, S.P. 2006. WikiRelate! Computing semantic relatedness using Wikipedia. American Association for Artificial Intelligence, 2006, Boston, MA.

Voss, J. 2006. Collaborative thesaurus tagging the Wikipedia way. Collaborative Web Tagging Workshop. Arxiv Computer Science eprints. URL: http://arxiv.org/-abs/cs/0604036

Zlatic, V., Bozicevic, M., Stefancic, H., and Domazet, M. 2006. Wikipedias: Collaborative web-based encylopedias as complex networks. Physical Review E, vol. 74