Profiling Topics on the Web

Aditya K. Sehgal
Department of Computer Science
The University of Iowa
Iowa City, IA 52242, USA
aditya-sehgal@uiowa.edu

Padmini Srinivasan
School of Library and Information Science &
Department of Computer Science
The University of Iowa
Iowa City, IA 52242, USA
padmini-srinivasan@uiowa.edu

ABSTRACT

The availability of large-scale data on the Web motivates the development of automatic algorithms to analyze topics and identify relationships between topics. Various approaches have been proposed in the literature. Most focus on specific entities, such as people, and not on topics in general. They are also less flexible in how they represent topics/entities.

In this paper we study existing methods as well as describe preliminary research on a different approach, based on profiles, for representing general topics. Topic profiles consist of different types of features. We compare different methods for building profiles and evaluate them in terms of their information content and ability to predict relationships between topics. Our results suggest that profiles derived from the full text present in multiple pages are the most informative and that profiles derived from multiple pages are significantly better at predicting topic relationships than profiles derived from single pages.

Keywords

text mining, web mining, profiles, information management

1. INTRODUCTION

The past two decades have seen the Web evolve from a specialized, closed-group medium to a very general, ubiquitous source of information. It is generally agreed that no source of information today compares to the Web in terms of sheer size and diversity. As it stands, the Web offers billions of pages containing information on almost every area that might be of interest to someone. The scale of the information available provides tremendous motivation for the development of automatic algorithms to "mine" the Web for interesting information. As a result, web mining has become a vibrant area of research.

A principal goal in web mining is to facilitate analysis of web data relevant to a topic of interest as well as to identify (or predict) relationships between topics. The existing literature offers many different approaches in this regard. An analysis reveals that most (e.g., [1, 13, 2]) focus on specific entities, mostly people entities. The focus is seldom on topics in general.

We define a topic as any subject of interest to a user. *Bill Clinton*, *A1BG Gene*, *Rainfall in the United States*, and *Cancer in Children* are all examples. Observe that while the first two are also entities, the latter are not. In general one can liken any web search query to a topic. Topics may also be identified by other types of text units such as by one or more sentences. Our interest is in web mining methods that are not constrained to particular varieties of topics. As an example, given an arbitrary group of topics, we would like to explore implicit links between them and identify new relationships.

Another observation that motivates our research is regarding the differences between various web mining approaches along two major dimensions as depicted in figure 1. The first dimension represents the number of pages used to represent a topic. This can range from a single page, such as a home page, to several hundred pages retrieved from a search engine. The second dimension refers to the level of data on a given page from which features descriptive of the topic are extracted. Two options are repeatedly used in the literature. The first designated the *instance* level uses text data including and surrounding individual mentions (instances) of the topic (more specifically entity) in a page. The second option, *page* level, extracts features from all of the page.

Figure 1 shows the different combinations of these two dimensions. The first quadrant (SPI) depicts approaches (e.g., [2]) that use instance-level data from a single page to derive features. The second quadrant (SPP) depicts approaches (e.g,. [1]) that use full text from a single page. The third quadrant (MPI) depicts approaches (e.g., [13]) that use instance-level data from multiple pages. Fourth quadrant (MPP) methods (e.g., [12]) use full text from multiple pages to derive features. Importantly, at this point the relative merits of these combinations are unknown.

Note that approaches based on SPI and MPI can only be applied to entities and not topics in general as it would be challenging to find complex topics explicitly represented by specific phrases. Also approaches based on SPP and SPI data utilize information in only a single page. Our own inclination is to explore methods from the MPP quadrant. This is because while a single page may contain information relevant to a topic, it is unlikely to contain all the relevant information. Topical information is likely scattered across multiple pages, each potentially addressing different relevant aspects. Moreover, it is quite possible for relevant sentences to appear distant from sentences that contain an instance of the topic. E.g., in our dataset a document relevant to the topic 'Hurricane Andrew' has the sentence *Hurricane Andrew was the most destructive United States hurricane of record*, which is relevant and contains an instance of the topic. But four sentences away there is: *The vast majority of*

WWW2007, May 8–12, 2007, Banff, Canada.
.

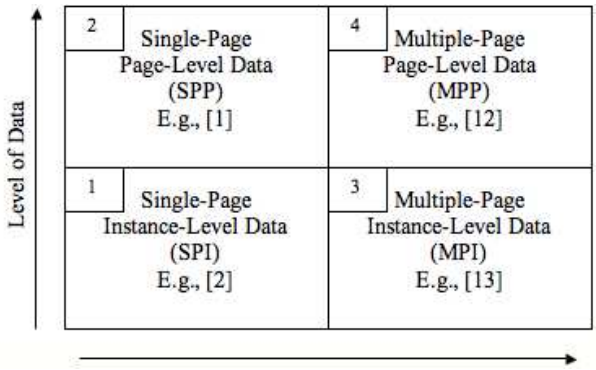

Figure 1: Different combinations of # of pages and level of data used, to generate representations. # of documents varies along horizontal axis and level of data varies across vertical axis.

the damage in Florida was due to the winds. This sentence is also relevant but does not contain an instance of the topic.

Another significant dimension that differentiates various web mining approaches relates to how topics/entites are represented. More specifically the difference is in the kinds of features used to represent topics. Example representations include features composed of words and named entities as in [12]. In [1] mailing lists subscribed to, words and links are features representing people entities. Although our current research is restricted to weighted word stems, links and named entities it is worth outlining the kind of representation we envisage in our research. Our long-term interest is in building extensible representations from the web that can accommodate features such as key concepts, relations and links. We have realized some of these interests in the context of topic profiles extracted from MEDLINE[1] using Manjal, our prototype biomedical text mining system [14]. An example for the web is given in Figure 3 with Bill Clinton as the topic. Formally, a topic profile is a composite vector of sub-vectors, each containing features of a particular type. Each feature is weighted to reflect its relative importance to the topic.

In summary, our observations regarding the characteristics of current web mining research and our own interests in topics beyond entities motivate our research. We seek a framework for automatically representing topics of all kinds using profiles and for analyzing relationships between them. Generally our *topical* perspective leads us to a view of a higher-level web, one where topics, represented by their profiles, and not pages are the fundamental unit of information. This is illustrated in figure 2. A key advantage is that relationships between topics can be analyzed independent of links between individual pages. Relationships may also be fine grained i.e., restricted to specific subsets of feature types. We believe that our higher-level topical web view is largely consistent with the "Entity-Centric Information and Knowledge Management" emphasis of the workshop. Additionally our topics include more abstract topics besides entities.

In the next section we briefly discuss related research. Following that we outline our methods for building SPI, SPP, MPI and MPP based profiles. Three different types of features are explored: words, links and named entities. We compare profile building methods through two experiments. First we use gold standards to evaluate the content of the profiles derived (section 4.1). Second we compare these profiles in their ability to predict relationships (section 4.2). We then analyze potential sources of error (section 5) and finally discuss our results and outline future steps.

[1]www.ncbi.nlm.nih.gov/entrez/

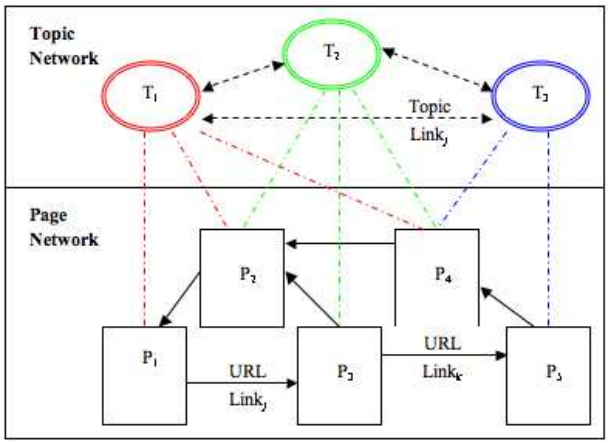

Figure 2: Topic-based web network vs. Page-based web network. In the upper image topics are unit of analysis and links are directly between topics. In the lower image pages are basic units and links are explicit hyperlinks.

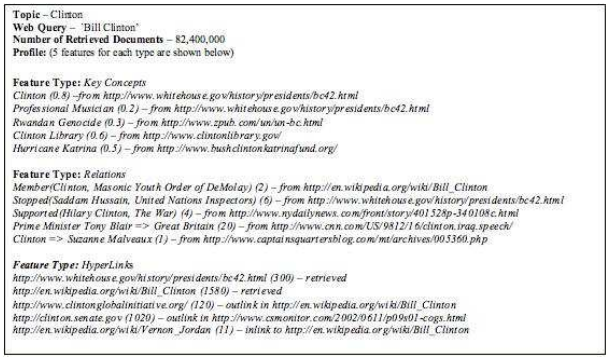

Figure 3: Example profile for the topic 'Bill Clinton' with 3 types of features.

2. RELATED RESEARCH

The problem of representing topics/entities using web data and predicting relationships between them is very well known in the web mining community. As mentioned before a number of approaches proposed in this regard can be found in the literature.

Most of the existing web-based research focuses on specific types of entities, mainly people entites. E.g., Ben-Dov et al. [2], Raghavan et al. [13], and Adamic and Adar [1]. Far less research [12] is seen with general topics. As mentioned in the introduction, most approaches fall under 4 categories, depending upon how many pages they use to derive representations and what data they use within a page.

Using a single page, such as a home page, is a standard

approach. For example, Adamic and Adar [1] represent students using data extracted from their home page. Ben-Dov et al. [2] use an instance within a single page to represent a person entity. Few approaches [3, 13] use data from multiple pages to represent entities. In [13] Raghavan et al. use instance-level data extracted from multiple pages to derive entity representations. Such efforts are risk missing potentially useful information outside these text windows. The example given earlier regarding the Hurricane topic illustrates this. A key disadvantage is also that such representations cannot be applied to general topics such as *Rainfall in the United States*. Relevant pages may not contain this phrase while still dealing with the topic. Even fewer approaches use full-text data from multiple web pages. In [12] Newman et al. represent topics and entities using words and named-entities, and associated topics, respectively. However, their choice of features is somewhat arbitrary and also their representations are fixed. In contrast, we explore approaches that exploit multiple pages, are not instance-based, and are able to accommodate a variety of features.

A number of approaches have been proposed to predict and analyze relationships between entities on the Web. Most rely on explicit indicators, such as shared hyperlinks [7, 4] or co-occurrence [6, 2, 11] to infer a relationship between two entities. Others [1, 13], while being independent of this requirement, are limited by the use of instance or page -level data to represent entities. Our topic profiles provide a framework for analyzing relationships between topics/entities based on various types of common features. Each type of feature provides a distinct thread that potentially binds two topics together. Consequently different forms of relationships can be analyzed between a pair of topics using our profiles. Also, profile-based relationships do not depend on shared hyperlinks or co-occurrence, which allows for analysis of more implicit relationships.

Interestingly, the literature reveals existing structures that are similar to ours but have been designed for different purposes or are limited in different aspects. Li et al. [8] create *entity profiles* limited to people, with only two types of features, viz., salient concepts, such as name and organization, and relationships to other entities (people). Glance et al. [5] generate high-level summaries of products. Features are limited to 4 measures generated using data from blogs and message boards. Liu et al. [10] generate descriptions of topics consisting of subtopics and their corresponding urls. This is analogous to topic profiles with only two types of features. Adamic and Adar [1] represent entities (students) using features, such as hyperlinks, text, and subscribed mailing lists, extracted from their home pages. Newman et al. [12] represent topics using words and named entity features extracted from multiple pages and entities using topic features. They generate social networks for entities based on the similarity of their representations.

While these are similar to our topic profiles, there exist substantial differences. Most representations are limited to specific types of entities, such as people [8, 1] or products [5]. These structures also consist of only specific kinds of features. Importantly, all the above efforts, except Adamic's and Newman's do not go beyond creating topic/entity synopses while we study topic representations in the context of using them for higher-level web mining applications.

3. METHODS

Our major goal is to compare different methods for building topic profiles. Following figure 1, we create various types of profiles differing in terms of the number of pages (Single or Multiple) and the level of data used (Instance or Page). Note that we use the same feature extraction strategy in each case.

We compare profiles in two different ways via two separate experiments. Firstly, we compare the quality of information in each type of representation and secondly we compare their ability to predict relationships between topics. In the first experiment we build different types of profiles for general topics using the SPI, SPP, MPI and MPP and compare them with profiles created from a known high quality source of information. Our topics are a mix of celebrities, important events and large corporations.[2]. In the second set of experiments we build the 4 types of profiles (SPI, SPP, MPI and MPP) for topics representing human proteins, and predict interactions between pairs of proteins based on how similar their profiles are. We evaluate the accuracy of these predictions using a high quality knowledge base of protein interactions.

The profile building process consists of two steps. In the first step we retrieve relevant pages using the Google search API and extract the title text, body text, anchor text, and hyperlinks from the individual pages. While this step mostly relies on the accuracy of the search engine to retrieve relevant pages, the retrieved pages may be further filtered (as is done in the first experiment; described in detail below). Multiple-page profiles are derived from the top N filtered set of pages while single-page profiles are derived from the home page. In case there is no home page then the top ranked filtered page is used. For instance-level profiles, the documents are further processed to eliminate sentences that do not contain an individual instance of the topic. We evaluated two sentence boundary detection tools, viz., Lingua EN[3] and MxTerminator[4] and use the latter as we found it to be more accurate through preliminary experimentation.

The second step consists of identifying the individual features that form part of the profiles and assigning each a weight. Three types of features are explored, words, links and named-entities, extracted from the title, body and anchor text. Individual words are stemmed and stop words are removed. Each feature is assigned a tf*idf weight. We describe the individual profiles in each experiment in more detail below.

4. EXPERIMENTS & RESULTS

4.1 Experiment 1: Information Content

In this experiment we evaluate topic profiles based on their information content. Specifically, we build SPI, SPP, MPI and MPP topic profiles and compare each with profiles for the same topics built from a known high quality source of information. We choose Wikipedia as our source of gold standard information. Wikipedia is an online repository of information on various topics in different languages. The

[2]Our choice of topics is influenced by the fact that SPI and MPI profiles can only be derived for entities and not general topics. Thus in order to compare all quadrants in figure 1, all the topics we choose are in fact entities.

[3]http://search.cpan.org/dist/Lingua-EN-Sentence/

[4]www.id.cbs.dk/ dh/corpus/tools/MXTERMINATOR.htm

English version of the site contains descriptions of more than a million topics (as of November 2006). A Wikipedia entry for a topic typically contains a summary, a small table listing prominent characteristics, a table of important relations, relevant external links, and references. A Wikipedia entry can be created by anyone and can also be edited by anyone. Thus, well-developed entries tend to contain the viewpoints of many people. Wikipedia is the largest collaborative journalism effort till date and is viewed as a highly regarded reference site [9]. It has been reported that the quality of Wikipedia articles is high and they are referenced by many teachers[5] and are also frequently cited by newspapers [15]. We also use the online version of the Encyclopedia Britannica (EB) as another source of high quality information. In contrast to Wikipedia, the data in EB is controlled, being manually compiled by trained personnel.

We compiled a list of 50 topics belonging to three categories, viz., famous people, important events of the 20th century and large companies. We randomly selected 16 people from the Forbes celebrity list for 2006 and 21 companies from the Fortune 500 list for the same year. The 'event' topics were randomly compiled from a list of 30 major 20th century events[6]. For each topic, we downloaded its Wikipedia page and derived profiles from word, link, and named-entity features extracted from the text. EB contained entries for only 26 of the 50 topics. We processed these pages in the same way.

For each topic, we manually identified relevant synonyms and generated boolean (OR) web search queries. These were used to retrieve the top 100 web pages for each topic using the Google Search API. The retrieved sets were filtered to exclude pages from approximately 600 web sites known to mirror Wikipedia content[7] (including Wikipedia itself). We then created the four different types of profiles and computed the cosine similarity between each profile and corresponding Wikipedia and EB profiles.

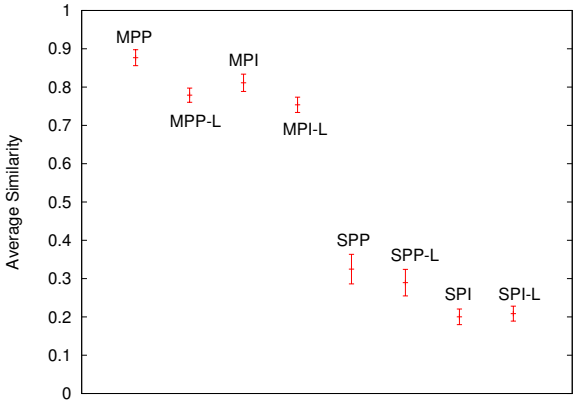

Figure 4: Performance of two variations of profiles against Wikipedia. (*-L) profiles contain link features in addition to word features in the other types of profiles.

Figure 4 shows the average similarity scores (along with 95% confidence intervals) of profiles gauged against Wikipedia profiles. Two types of profiles are created for each approach. Those derived from word features extracted from the title, body, and anchor text (MPP, MPI, SPP, SPI), and those additionally containing link features (MPP-L, MPI-L, SPP-L, SPI-L). In the link-inclusive profiles (*-L), text information was assumed to be more important and thus assigned a higher weight (0.9) than link information (0.1).

First we see that profiles derived from multiple pages are significantly better (statistically at 0.05 level) than single-page profiles. For MPP the difference between both variations is also significant, with the non link version being better. But the same is not the case for the other three types of profiles. For these link features are not useful. Assigning a lower weight to links did not change this observation. Consequently, in subsequent experiments, we exclude link information from profiles. Interestingly, the average similarity for MPP profiles is quite high (0.88). A key observation to note is that MPP profiles are significantly better than MPI profiles.

Figure 5 compares MPP and MPI profiles derived from varying numbers of relevant pages, with Wikipedia as the gauge. Interestingly the difference between the two types of profiles decreases as the number of pages increases. At each point, however, MPP is statistically better than MPI at the 0.05 level. We note that the average score for MPP profiles stabilizes at 10 pages and for MPI profiles at 15 pages.

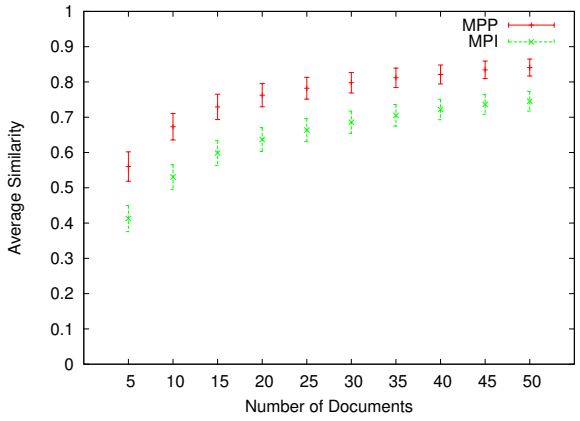

Figure 5: Performance of profiles by varying numbers of documents, against Wikipedia.

Figure 6 compares the different profiles with corresponding Wikipedia (WK subscript) and EB profiles (EB subscript). We see that average scores are much higher when Wikipedia is the gauge than EB. Also the variation in scores is lower. Hence we have greater confidence in our Wikipedia-based observations. We find no significant difference between MPP_{EB} and MPI_{EB}. But the two are again significantly better than page-based profiles.

This figure also reveals interesting (and significant) differences between using Wikipedia and EB as gold standards. E.g., there is a consistent difference in length of the confidence bars. In an effort to understand why this was so, we closely analyzed the EB entries, which revealed a significant variation in the amount of text they contain. Some entries are very long (e.g., over 50 pages for *WWII*) while others are very small (e.g., only a small paragraph for *Bank of Amer-*

[5]http://meta.wikimedia.org/wiki/Research
[6]http://history1900s.about.com/cs/majorevents/
[7]http://en.wikipedia.org/wiki/Wikipedia:Mirrors_and_forks

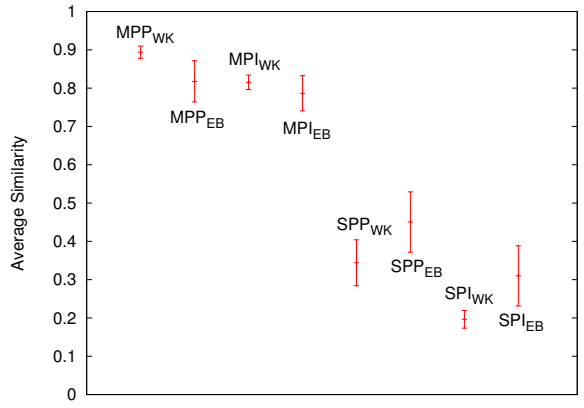

Figure 6: Performance of profiles against Wikipedia and EB.

ica). Also, most EB entries are quite small, in comparison to corresponding Wikipedia entries. We believe that the average scores for different profiles, particularly those derived from multiple pages (MPP and MPI), are significantly affected by the smaller size of EB entries. To confirm this hypothesis, we segregated the topics into two groups, those with EB entries less than 5 KB (14 topics) and those with EB entries greater than or equal to 5 KB (12 topics). We then repeated our analysis for these two groups.

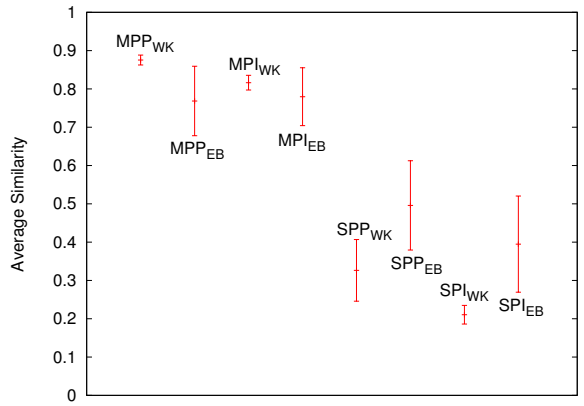

Figure 7: Performance of profiles for topics with EB entries less than 5 KB in size.

Figures 7 and 8 show that average scores are generally higher for topics with larger EB entries. The confidence intervals are considerably smaller. Note that the Wikipedia results are also shown for the two topic subsets. The difference between MPP_{EB} and MPI_{EB} profiles is also significant at the 0.05 level for topics with larger EB entries while this is not the case for topics with smaller EB entries. But in both topic subsets the average scores with Wikipedia are much higher than with EB. This may be because, in general, Wikipedia entries have more text than corresponding EB entries. We also see that profiles derived from multiple pages are always significantly better than profiles derived from single pages.

Figure 9 shows the average scores for MPP and MPI pro-

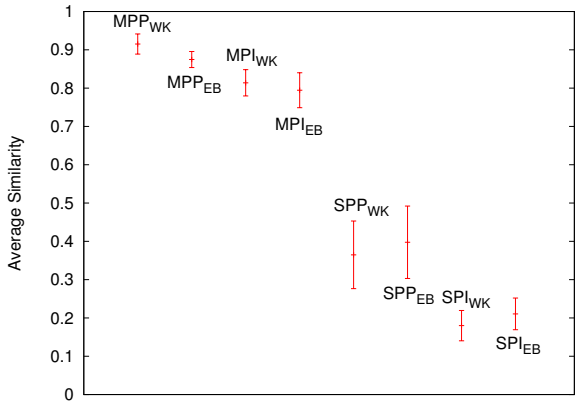

Figure 8: Performance of profiles for topics with EB entries greater than 5 KB in size.

files derived from varying numbers of relevant pages, with EB as the gold standard. As was the case in figure 5, the difference between their average similarities decreases as the number of pages increases. However, in this case the decrease is much more rapid. These differences are significant for profiles derived from 20 or less relevant pages. We also note that the average similarity score stabilizes at 10 pages for both MPP and MPI.

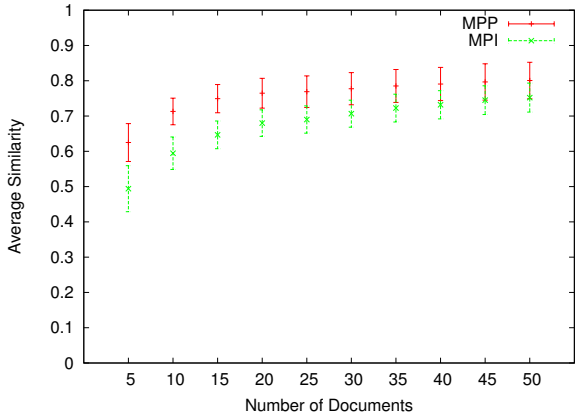

Figure 9: Performance of profiles by varying numbers of documents, against EB.

Figure 10 offers the same analysis but for topics with EB entries greater than 5 KB. Here again we see the difference between the average scores decreasing with increasing number of pages but not as rapidly as in figure 9. Here MPP and MPI profiles are significantly different for 30 or less pages. Also, the average score stabilizes at 15 pages for both types of profiles.

We now take a different approach and extract named-entity features. Figure 11 shows the average similarity scores for profiles with such features. Three types of named entities, Person, Location, Organization, were extracted from retrieved web pages. We tried two named-entity recognitions tools, viz., Stanford-NER[8] and Lingpipe and found

[8]http://nlp.stanford.edu/software/CRF-NER.shtml

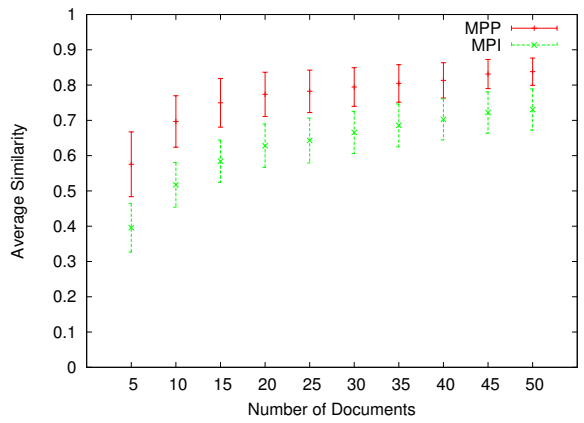

Figure 10: Performance of profiles for topics with EB entries greater than 5 KB by varying numbers of documents, against EB.

the former to be more accurate through preliminary experimentation. Hence we use the Stanford-NER system. We see that while MPP profiles have the highest average similarity, the difference between them and MPI profiles is not statistically significant. However, profiles derived from multiple pages continue to be significantly better than single-page profiles. We note a considerable drop in average similarity scores compared with word-based profiles (e.g., from 0.88 to 0.6 for MPP).

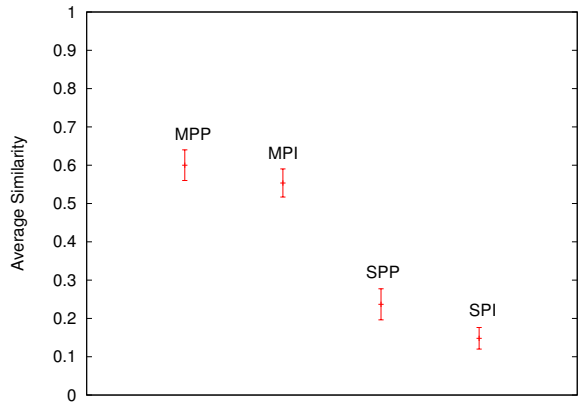

Figure 11: Performance of profiles with named-entity features against Wikipedia.

4.2 Experiment 2: Prediction Ability

In the second set of experiments we evaluate the different types of profiles based on their ability to predict known relationships between proteins. We randomly compiled a list of 82 proteins from the Database of Interacting Proteins (DIP)[9]. According to DIP, 90 pairs of proteins within this set are known to interact with each other. This forms our positive gold standard set of interactions. The remaining pairs are considered as the negative set of interactions. For each protein, we identified synonyms from the SwissProt protein database[10] and excluded those that are English words (and thus cause for ambiguity). We then created web search queries from these and for each query downloaded the top 100 pages using the Google API. For each protein topic, we build the various types of profiles (MPP, MPI, SPP, and SPI) and compute similarities between pairs of profiles. Two types of features were extracted from documents, word features and named-entities, from the title, body and anchor text. As before each feature is assigned a tf*idf weight.

We predict relationships between proteins based on the similarity of their profiles. We use the standard IR cosine similarity score to measure similarity between profiles. Two proteins are predicted to be related if their similarity score is above a certain threshold. Based on predictions made we measure precision, recall and f-score. In our experiments we use f-score[11] as the primary measure to evaluate different types of profiles.

We adopt a 5-fold cross-validation approach. The 82 proteins we selected yielded 3403 unique pairs[12]. These were randomly split into five sets of equal size, each with the same ratio of positives and negatives. We consider the union of four of the splits (*folds* in standard machine learning parlance) as our training set and the remaining split as our test set. We choose a threshold that optimizes the training f-score and then apply this threshold to the test set and compute the test f-score. This process is repeated five times, each time with different combinations of splits considered as the training data. Different profiling building methods are compared based on the average of their five test f-scores.

For our first experiment we created profiles consisting of word features found in the title, body and anchor text of pages. As before, we consider four types of profiles (MPP, MPI, SPP, and SPI). Table 1 shows the average training and testing precision, recall and f-scores for each type of profile across all five iterations. We see that the average test f-scores for MPP, MPI and SPI (0.25) are very close and thus they are similar in their ability to predict relationships. However, all three are significantly (at 0.05 level) better than SPP profiles. Also the training f-scores are in general similar to the testing f-scores, suggesting that the thresholds computed using the training data were general enough to apply to new (test) data.

Figure 12 shows the average f-scores for MPP and MPI profiles when the number of relevant pages used to build the profiles are varied. Here again profiles consist of word features extracted from the title, body and anchor text of relevant pages. We see that varying the number of pages has little affect on the average f-scores for both types of profiles and in general we see no significant difference in performance between the two. We also note that the average f-score for both stabilizes at 5 documents.

We next created profiles that contained more specific features, viz., named-entities. Named entities were extracted from relevant pages using the lingpipe[13] named entity extraction package. This package comes with a model pretrained on the GENIA[14] corpus, which is what we used for this experiment. Figure 13 shows the average f-scores for

[9]http://dip.doe-mbi.ucla.edu/

[10]http://expasy.org/sprot/

[11]with equal weight to precision and recall

[12]taking into account symmetry so that only P1->P2 is considered and P2->P1 is not

[13]www.alias-i.com/lingpipe/

[14]www-tsujii.is.s.u-tokyo.ac.jp/GENIA/

Type	Split	Training			Testing				
		Precision	Recall	Fscore	Precision	Recall	Fscore	LCI (95%)	UCI (95%)
MPP	Average	0.2205	0.3528	0.2704	0.2046	0.3333	0.2529	0.1826	0.3232
MPI	Average	0.2188	0.3222	0.2604	0.2115	0.3111	0.2479	0.2061	0.2897
SPP	Average	0.1183	0.2702	0.1628	0.1011	0.2303	0.1357	0.0963	0.1751
SPI	Average	0.2190	0.3222	0.2607	0.2101	0.3111	0.2501	0.1885	0.3117

Table 1: Average training and test precision, recall and f-scores for the four types of profiles. LCI and UCI denote lower and upper 95% confidence intervals of test f-score, respectively.

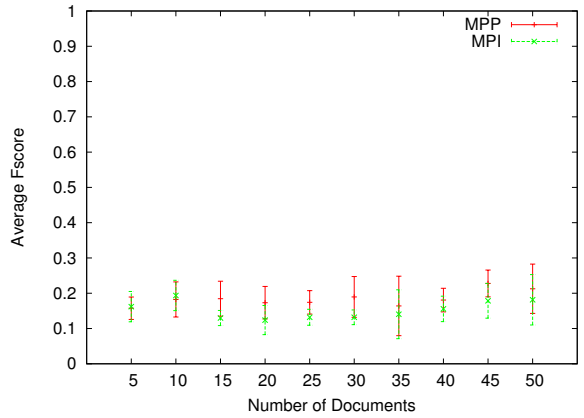

Figure 12: Average prediction f-scores for profiles by varying numbers of pages.

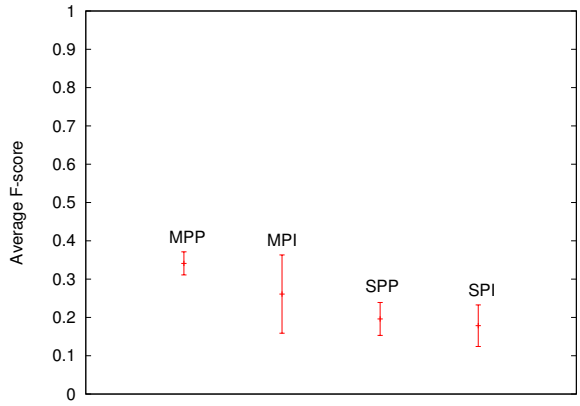

Figure 13: Average prediction f-scores for profiles with named-entity features.

the four different types of profiles. Here again, we see no significant difference between MPP and MPI profiles. However, MPP profiles are significantly better than single-page profiles. Comparing named-entity profiles with word profiles, we see that the average MPP, MPI, and SPP f-scores are higher for the former than corresponding f-scores for the latter (shown in table 1), while the reverse is true for SPI.

5. POTENTIAL SOURCES OF ERRORS

Our approach relies upon several underlying technologies, such as document retrieval, sentence detection, and named-entity recognition. Each of these is a potential source of error. First, we rely on the accuracy of search engines to retrieve relevant documents for topics, from which profiles are derived. While modern search engines are fairly accurate, they are still vulnerable to problems such as ambiguity. Despite our attempts, some ambiguity still remained. E.g., 'NEMO' is a synonym for a protein and is also the name of a popular cartoon character.

In this paper we have created profiles from sentences in relevant documents containing individual instances of the topics. Sentence boundary detection in web pages is a hard problem and many off-the-shelf tools (we use MxTerminator) are not optimized for web pages. Due to the tag structure of web pages, many incoherent sentences are identified.

In this research we have also evaluated profiles containing named-entity features extracted from relevant pages. Named-entity recognition is a challenging problem and primarily depends on the training data used to train the model. Our use of models pre-trained on newswire data in the first experiment and GENIA data in the second resulted in some errors, due to the fact that there are significant differences between general web text and newswire and GENIA text. E.g., EBI (European Bioinformatics Institute) was recognized as a protein by Lingpipe.

6. DISCUSSION

Our experiments have yielded interesting results. In section 4.1 our results clearly show that profiles derived from full text in multiple pages are more informative than single-page profiles. They are also better than profiles derived from multiple pages but restricted to instance-level text windows. This is when Wikipedia is used as the gold standard. It is also true with EB as the gold standard for topics that have entries that are at least 5 KB in size. These results support our intuition that relevant information tends to be scattered over multiple pages and is not necessarily instance bound.

Our results also suggest that adding the link information present in relevant pages as features does not improve the information content of a profile. In fact assigning a higher weight to links had a detrimental effect. However, this could be attributed to the relatively small number of links present in Wikipedia pages, which would bring down the average similarity score.

The difference between MPP and MPI profiles was much less prominent when they contained specific features, such as named entities. The average similarity of such profiles w.r.t. Wikipedia was also notably lower than for corresponding profiles containing word features.

Also, we found that the difference between MPP and MPI was greatest when few relevant pages were present. As the number of relevant pages increased, the difference between

the two shrank. This suggests that when few relevant pages are available, as is quite often the case, it would be better to use the full-text to create representations. A key observation made is that in general 10 to 15 pages are sufficient for MPP, our best strategy, to achieve its highest similarity with the gold standard.

As an aside our experiments also revealed interesting differences between the two different gold standard sources of information we considered, Wikipedia and EB. In general we found the former to be more comprehensive for the topics we selected, while the latter contained comprehensive information only for a few popular topics, such as WWI and WWII, and cursory information for the rest.

While significant differences between MPP and MPI profiles were detected in the first set of experiments, this was not the case with the second set of DIP experiments testing prediction ability. In general, irrespective of the type of features contained in the profile, the performance of MPP and MPI profiles in predicting known DIP relationships was very similar. However, MPP profiles have the important advantage that they can model general topics, while MPI profiles can only be built for entities. Thus it would be preferable to use the former. Again protein profiles derived from multiple pages were always significantly better than profiles derived from single pages, confirming our intuition that the latter did not contain enough information for a good representation. We also found that varying the number of relevant pages did not have any affect on the prediction ability of MPP and MPI profiles.

As a final point, our DIP based experiments also let us test the idea of implementing a bioinformatics web mining application. Although the results are moderate these offer a reasonable starting point. In future work we will also compare these results with similar experiments run against MEDLINE. Given the vastness of the Web, it is fast becoming a data and knowledge source for domain specific applications.

In conclusion, this research contributes to our long-term goal which is to be able to represent arbitrary topics on the web with topic profiles consisting of weighted features of different types. We see value in pursuing a higher level topical web where the object (node) of interest is the topic and the link represents inter-topic relationships. Such a web has the potential to more effectively support individual information needs as well web mining applications seeking to discover novel connections between topics.

7. ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation under Grant No. 0312356 awarded to Padmini Srinivasan. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

8. REFERENCES

[1] L. Adamic and E. Adar. Friends and Neighbors on the Web. *Social Networks*, 25(3):211–230, 2001.

[2] M. Ben-Dov, W. Wu, P. Cairns, and R. Feldman. Improving knowledge discovery by combining text-mining and link analysis techniques. In *Proceedings of the SIAM International Conference on Data Mining*, 2004.

[3] J. Conrad and M. Utt. A system for discovering relationships by feature extraction from text databases. In *Proceedings of the 17th Annual International ACM-SIGIR Conference (Special Issue of the SIGIR Forum)*, pages 260–270, 1994.

[4] G. Flake, S. Lawrence, and C. Giles. Efficient identification of web communities. In *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 150–160, 2000.

[5] N. Glance, M. Hurst, K. Nigam, M. Siegler, R. Stockton, and T. Tomokiyo. Deriving marketing intelligence from online discussion. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2005)*, pages 419–428, 2005.

[6] H. Kautz, B. Selman, and M. Shah. Referral Web: Combining social networks and collaborative filtering. *Communications of the ACM*, 40(3):63–65, 1997.

[7] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the web for emerging cyber-communities. In *Proceedings of the 8th International World Wide Web Conference (WWW-1999)*, pages 403–415, 1999.

[8] W. Li, R. Srihari, C. Niu, and X. Li. Entity profile extraction from large corpora. In *Proceedings of Pacific Association for Computational Linguistics 2003 (PACLING-2003)*, 2003.

[9] A. Lih. Wikipedia as participatory journalism: Reliable sources? metrics for evaluating collaborative media as a news resource. In *Proceedings of the 5th International Symposium on Online Journalism*, 2004.

[10] B. Liu, C. Chin, and H. Ng. Mining topic-specific concepts and definitions on the web. In *Proceedings of the twelfth international World Wide Web conference (WWW-2003)*, pages 251–260, 2003.

[11] Y. Matsuo, J. Mori, M. Hamasaki, K. Ishida, T. Nishimura, H. Takeda, K. Hasida, and M. Ishizuka. Polyphonet: an advanced social network extraction system from the web. In *Proceedings of the 15th International World Wide Web Conference (WWW-2006)*, pages 397–406, 2006.

[12] D. Newman, C. Chemudugunta, P. Smyth, and M. Steyvers. Analyzing entities and topics in news articles using statistical topic models. In *IEEE International Conference on Intelligence and Security Informatics*, pages 93–104, 2006.

[13] H. Raghavan, J. Allan, and A. McCallum. An Exploration of Entity Models, Collective Classification and Relation Description. In *Proceedings of KDD Workshop on Link Analysis and Group Detection*, 2004.

[14] A. Sehgal and P. Srinivasan. Manjal - A Text Mining System for MEDLINE (demonstration). In *Proceedings of the 28th Annual International ACM SIGIR*, page 680, 2005.

[15] J. Voss. Measuring wikipedia. In *Proceedings of the 10th International Conference of the International Society for Scientometrics and Informetrics*, 2005.