SEMANTIC TAGGING FOR DOCUMENTS USING 'SHORT TEXT' INFORMATION

Ayush Singhal[1] and Jaideep Srivastava[1]

[1]Department of Computer Science & Engineering,
University of Minnesota, Minnesota, USA
singh196,srivasta@umn.edu

ABSTRACT

Tagging documents with relevant and comprehensive keywords offer invaluable assistance to the readers to quickly overview any document. With the ever increasing volume and variety of the documents published on the internet, the interest in developing newer and successful techniques for annotating (tagging) documents is also increasing. However, an interesting challenge in document tagging occurs when the full content of the document is not readily accessible. In such a scenario, techniques which use "short text", e.g., a document title, a news article headline, to annotate the entire article are particularly useful. In this paper, we propose a novel approach to automatically tag documents with relevant tags or key-phrases using only "short text" information from the documents. We employ crowd-sourced knowledge from Wikipedia, Dbpedia, Freebase, Yago and similar open source knowledge bases to generate semantically relevant tags for the document. Using the intelligence from the open web, we prune out tags that create ambiguity in or "topic drift" from the main topic of our query document. We have used real world dataset from a corpus of research articles to annotate 50 research articles. As a baseline, we used the full text information from the document to generate tags. The proposed and the baseline approach were compared using the author assigned keywords for the documents as the ground truth information. We found that the tags generated using proposed approach are better than using the baseline in terms of overlap with the ground truth tags measured via Jaccard index (0.058 vs. 0.044). In terms of computational efficiency, the proposed approach is at least 3 times faster than the baseline approach. Finally, we qualitatively analyse the quality of the predicted tags for a few samples in the test corpus. The evaluation shows the effectiveness of the proposed approach both in terms of quality of tags generated and the computational time.

KEYWORDS

Semantic annotation, open source knowledge, wisdom of crowds, tagging.

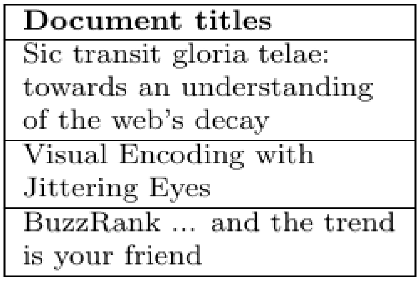
1. INTRODUCTION

Tagging documents with relevant and comprehensive keywords offer an invaluable assistance to the readers to quickly overview any document [20]. With the ever increasing volume and variety of the documents published on the internet [12], the interest in developing newer and successful techniques for tagging documents is also increasing. Tagging documents with minimum words/key-phrases have become important for several practical applications like search engines, indexing of databases of research documents, comparing the similarity of documents, ontology creation and mapping and in several other stages of important applications [4]. Although

Dhinaharan Nagamalai et al. (Eds) : ACITY, WiMoN, CSIA, AIAA, DPPR, NECO, InWeS - 2014
pp. 337–350, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4534

document tagging is a well-studied problem in the field of text mining, but there are several scenarios that have not drawn sufficient attention from the scientific community.

Table 1: A few examples of document titles which do not try to capture the essence of the document's content.



Document titles
Sic transit gloria telae: towards an understanding of the web's decay
Visual Encoding with Jittering Eyes
BuzzRank ... and the trend is your friend

A few of the challenges regarding document tagging, which is not well addressed in the literature are: (1) entire content of the document is not accessible due to privacy or protection issues (2) document heading does not summarize the content of the document (3) reading entire document is time consuming. The first challenge requires techniques to generate tags using only a short description of the document (document heading/title, snippet). The second challenge requires ’intelligence’ to figure out the context represented by the heading or title. As an example, consider the examples shown in table 1. This table shows a few examples of ‘catchy’ titles used in scientific research articles to provide headings of the articles. Only using such title information it would be hard to delve into the subject matter of these articles. The third challenge is particularly relevant in situations when the document itself is quite large and thus, requires tagging using only partial information from the document for quick annotation. The third challenge is particularly relevant in the case of real-time response systems.

To the best of our knowledge, the above mentioned challenges have not been well addressed in the literature. Most of the current literature provide efficient techniques for key- word extraction using the text content from single or multiple documents [17, 8]. Such techniques are not suitable if the text content of the document is very short or unavailable. Another class of problems which is of increasing interest is that of key-phrase abstraction. While these techniques do not extract keywords directly from the text content, the text content is required to build models for keyword abstraction [7]. However, the area of keyword extraction is still in the developing stage. Eventually, the overall goal of these research directions is to automate the annotation of documents with key phrases that are very close to what a human could generate. We further elaborate upon the specific research works and milestones in section 7.

In this work we propose a novel approach to address the above mentioned challenges. We propose a novel approach that takes as input only a ’short text’ from the query document and leverages intelligence from the Web2.0 to expand the context of the ’short text’. We have used academic search engines to expand the context of the ’short text’. The expanded context utilizes the intelligence of the web to find relevant documents to overcome the ’catchiness’ of the title. The tags are generated using the world knowledge from DBpedia, Freebase, Yago and other similar open source crowd-sourced databases. Moreover, using the crowd- sourced knowledge bases ensures that the tags are up-to-date as well as popular. Finally, we propose an unsupervised algorithm to automatically eliminate the ’noisy’ tags. The un- supervised approach uses web-based distances (also famous as ’wisdom of crowd’ [10]) to detect outlier tags. The overall framework is fully unsupervised and therefore, suitable for real-time applications for any kind of documents.

In order to demonstrate the effectiveness of the proposed approach in real-world applications, we have used a sample of the dataset from the DBLP digital archive of computer science research articles [15]. We evaluate the performance of the proposed approach using 50 test documents. We also compare the performance of the proposed approach with a baseline approach which uses the full content of the documents in order to generate tags. Surprisingly, we find that the tags generated by the proposed approach, which uses only the title/heading information of the document to predict tags, has a greater overlap (measured via the Jaccard index) with the ground truth tags (0.058 vs. 0.044) in comparison to the baseline. We also find that the proposed approach is computationally at least 3 times faster than the baseline approach. A qualitative analysis of the generated tags for a few sample test documents further reveal the effectiveness of the proposed approach for semantic tagging using only 'short text' information from the document. Although the proposed approach is tested only on the DBLP dataset, the approach, however, is generic enough to be used for various types of documents like news articles, patents and other large content documents.

We have made the following contributions in this work:

- A novel approach for using 'short text' for context expansion using web intelligence.
- A novel approach for tag generation using crowd-sourced knowledge.
- A novel approach to eliminate 'noisy' tags using web- based distance clustering.
- We provide a quantitative and a qualitative validation on a real world dataset.

The rest of the paper is organized in the following manner. In Section 2, we define the problem statement. In section 3, we mention some of the important features of crowd-sourced knowledge bases. The details of the proposed approach are discussed in Section 4. Experimental design and the results are discussed in Section 5 and 6. Section 7 describes the related literature. Finally, the summary of the work and a few directions for future research are presented in Section 8.

2. PROBLEM FORMULATION

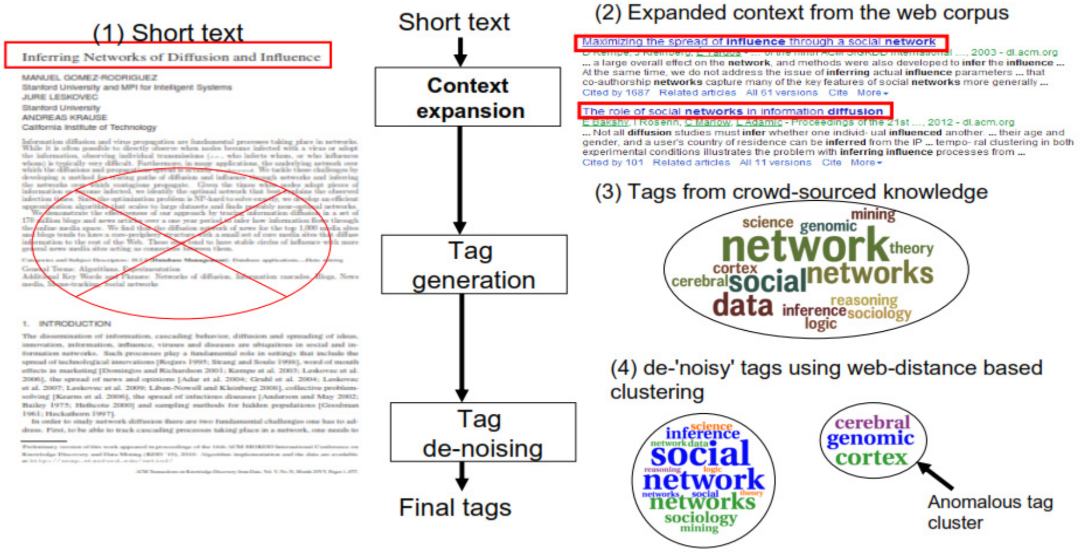
The problem of document tagging is formulated in the following manner. Given a document's text content S, the research problem is to identify k keywords/phrases based on the content S of the given document. In this case $k \ll$ size (S).

In the present work, we are studying a slightly different problem from the one describe above. We consider the title of the document as the only available text content (S'). We call this information about the document as the 'short text' since it is only a short description of the document. Also the size (S') $\ll$ size (S). The research problem is to find k keywords/phrases to describe the main topics/themes of the document. The keywords/phrases may not be directly present in the content of the document. Here, k is not known a priori.

3. BACKGROUND OF OPEN SOURCE KNOWLEDGE BASES

3.1 Crowd-sourced knowledge

Wikipedia is currently the most popular free-content, online encyclopedia containing over 4 million English articles since 2001. At present Wikipedia has a base of about 19 million registered users, including over 1400 administrators. Wikipedia is written collaboratively by largely anonymous internet volunteers. There are about 77,000 active contributors working on the articles in Wikipedia. Thus the knowledge presented in the articles over the Wiki are convinced upon by editors of similar interest.



Maximizing the spread of influence through a social network
D Kempe, J Kleinberg, É Tardos - … of the ninth ACM SIGKDD international …, 2003 - dl.acm.org
… a large overall effect on the network, and methods were also developed to infer the influence … At the same time, we do not address the issue of inferring actual influence parameters … that co-authorship networks capture many of the key features of social networks more generally …
Cited by 1687 Related articles All 61 versions Cite More

The role of social networks in information diffusion
E Bakshy, I Rosenn, C Marlow, L Adamic - Proceedings of the 21st …, 2012 - dl.acm.org
… Not all diffusion studies must infer whether one individ- ual influenced another. … their age and gender, and a user's country of residence can be inferred from the IP … tempo- ral clustering in both experimental conditions illustrates the problem with inferring influence processes from …
Cited by 101 Related articles All 11 versions Cite More

Figure 1. A systematic framework of the proposed approach. An example is illustrated to explain the proposed approach

DBpedia is another crowd-sourced community effort to extract structured information from Wikipedia and to make this information available on the Web. DBpedia allows you to ask sophisticated queries against Wikipedia, and to link the different data sets on the Web to Wikipedia data. The English version of the DBpedia knowledge base currently describes 4.0 million things, out of which 3.22 million are classified in a consistent ontology. For example DBpedia knowledge base allows you to ask quite surprising queries against Wikipedia, for instance “Give me all cities in New Jersey with more than 10,000 inhabitants” or “Give me all Italian musicians from the 18th century”.

Yago is similar to DBpedia. In addition to Wikipedia, Yago combines the clean taxonomy of WordNet. Currently, YAGO2s has knowledge of more than 10 million entities (like persons, organizations, cities, etc.) and contains more than 120 million facts about these entities. Moreover, YAGO is an ontology that is anchored in time and space as it attaches a temporal dimension and a spatial dimension to many of its facts and entities proving a confirmed accuracy of 95%.

Freebase is another online collection of structured data collected from various sources such as Wikipedia, ChefMoz, and MusicBrainz, as well as individually contributed user information. Its database infrastructure uses a graph model to represent the knowledge. This means that instead of using tables and keys to define data structures, its data structure is defined as a set of nodes and a set of links that establish relationships between the nodes. Due to its non-hierarchical data structure, complex relationships can be modeled be- tween individual entities.

3.2 Academic search engines

Academic search engines provide a universal collection of research documents. Search engines such as Google scholar and similar other academic search engines have made the task of finding relevant documents for a topic of interest very fast and efficient. We use the capacity of search engines to find relevant documents for a given query document. We have used the Google search engine for this purpose.

4. PROPOSED APPROACH

In this section, we discuss the framework of the proposed approach. The approach consists of three main components which will be discussed in detail in this section. The overall framework is summarized in the schematic (figure 1). As shown in this figure, the proposed approach for semantic annotation of a document is a three step procedure: (1) Context expansion using academic search engine, (2) candidate tag generation using crowd-sourced knowledge and (3) de-noising tags using web-based distance (a.ka. 'wisdom of crowd') clustering technique. Given a document as a short text S', the final results are k semantic tags, where k is not fixed apriori.

4.1 Context expansion

As mentioned earlier, the problem of reading the entire text content of the document or the lack of availability of the full text content restricts the task of tagging based on the document's text content. Moreover, techniques utilizing the text content of the document generate tags or keywords only from within the document's text content. While such key- word extraction approaches are necessary, but this might often restrict the keyword usage for the document. In such a scenario, it is helpful to generate keywords that are more popular and widely accepted for reference. To accomplish this goal, we propose a web-based approach to generate an expanded context of a document.

Given the 'short text' S information of a document, the expanded context is generated by mining intelligence from the web using an academic search engine. As shown in figure 1, the context of the 'short text' (S') is expanded using the results obtained by querying the web corpus with an academic search engine. The 'short text' is used as a query for the search engine. The new context of the 'short text' include the titles/heading (h) of the top-k results returned by the search engine. It is also possible to use other contents of the results like the snippets, author names, URLs to create an extended context. However, for this work, the approach is kept generic such that it is applicable to all sorts of search engines. The value of k is not fixed and can be a parameter to the approach. In the later section, the results are evaluated by varying the value of k.

The extracted results headings (h) that form the expanded context of the 'short text' are transformed into a bag of words representation. As a basic step in text mining, the bag of words is pre-processed by applying stop-word removal, non-alphabetic character removal and length-2 word removal techniques. In the rest of the paper, the expanded context of S is referred as C (S') for the sake of convenience and consistency. The final context created using the search engine is expected to contain a wider variety.

4.2 Tag generation

In this section, we describe the procedure to utilize crowd- sourced knowledge to generate tags from the expanded context C (S'). As described earlier, the crowd-sourced knowledge is available in well-structured format unlike the un- structured web. The structured nature of knowledge from sources such as DBpedia, Freebase, Yago, Cyc provides opportunity to tap in the world knowledge from these sources. The knowledge of these sources is used in the form of concepts and named entity information present in them, since the concepts and named entities consists of generic terms useful for tagging. We have used the AlchemyAPI [1] to access these knowledge bases. A tool such as this provide a one-stroke access to all these knowledge bases at once and returns a union of the results from all the various sources.

Given the expanded context (C (S')) as the input to the AlchemyAPI, which matches the C (S') against the indices of these knowledge sources to match C (S'), using the word frequency distribution, with concepts and named entities stored in the knowledge bases. The output for an API query C (S’) is a list of concepts and named entities. Using the open source knowledge bases and the word frequency information from the input, the API returns a list of concepts related to the content. The named entity list returned from a query C (S’) consist of only those named entities of type ’field terminologies’. There are other types of named entities such as ’person's name’, ’job title’, ’institution’ and a few other categories but those are not generic enough to be used as tags. The concepts and named entities for C (S’) together form a tag cloud T.

Figure 1 highlights a tag cloud consisting of tags generated using the above described technique. As shown in the figure, the tags are weighted based on the word distribution in C (S’). This example also shows a few tags like ’cerebral’, ’cortex’, ’genomic’ that appear to be inconsistent with the overall theme of the tag cloud for C (S). The next step describes an algorithm to handle such situations in the tagging process.

4.3 Tag cloud de-noising

As described in the previous step, the tag cloud T for C (S’) may contain some inconsistent or ’noise’ tags in it. In this section, we describe an algorithmic approach to automatically identify and prune ’noisy’ tags in the tag/keyword cloud. This step is therefore termed as tag cloud de-noising.

Given the tag cloud T for C (S’), noisy tags are pruned in the following manner. The tags in T are clustered using a pairwise semantic distance measure. Between any two tags in T, the semantic distance is computed using the unstructured web in the following way. For any two tags t_1 and t_2 in T, *dis* (t_1, t_2) is defined as the normalized Google distance (NGD) [2]:

$$NGD(t_1, t_2) = \frac{max\{logf(t_1),logf(t_2)\}-logf(t_1,t_2)}{logM-min\{logf(t_1),logf(t_2)\}}$$

where M is the total number of web pages indexed by the search engine: $f(t_1)$ and $f(t_2)$ are the number of hits for search terms t_1 and t_2, respectively; and $f(t_1,t_2)$ is the number of web pages on which both t_1 and t_2 occur simultaneously.

Using the NGD metric, a pairwise distance matrix (M) is generated for the tag cloud T. The pairwise matrix M is used to identify clusters in the tag cloud. Finally, the tag cloud is partitioned into two clusters using hierarchical clustering techniques. Here, we assume that there is at least one ’noise’ tag in the tag cloud T. Out of the two clusters identified from the tag cloud T, the one cluster with majority tags is called a normal cluster, whereas the other cluster is called as an outlier cluster (or noisy cluster). In case of no clear majority the tie is broken randomly.

The algorithm is illustrated through an example shown in figure 1 step 4. This step shows that the tags in the tag cloud T generated in step 3 are partitioned into two clusters as described above. The tags in one cluster are semantically closer than the tags in the other clusters, as per the ’wisdom of crowd’ semantics. As shown in this example, the outlier tags ’cerebral’, ’cortex’, ’genomic’ are clustered together while the remaining normal tags cluster together. Since the former is a smaller cluster, it is pruned out from the tag cloud. Lastly, the final tag cloud consisting only the larger cluster of tags is returned as the output.

5. EXPERIMENT ANALYSIS

This section describes the experimental design used to evaluate the performance of the proposed approach. In this section, we discuss the test dataset used for evaluation, the ground truth for evaluation, the baseline and the evaluation metrics used for evaluation of the proposed approach.

5.1 Test dataset description

For the purpose of evaluating our approach we use a test set consisting of 50 research documents from top tier computer science conferences constructed from the DBLP corpus [15]. The 50 papers were selected to capture the variety of documents in computer science research. Several of the documents had catchy titles (examples given in Table 1). The test data are accessible here (https://www.dropbox.com/sh/iqnynrixsh2oouz/8dWnbXhh7B?n=62599451).

5.2 Ground truth

In the absence of any gold standard annotations for the test documents, the ground truth of the documents was collected by the author assigned keywords to these documents. We collected this information by parsing these documents. We assume that the keywords assign by the authors are representative of the annotations for the document. The proposed approach and the baseline were evaluated on this ground truth.

5.3 Baseline approach

We have compared the performance of the proposed approach with a baseline, which uses the full text content of the test documents. In order to evaluate the claim that the ’short text’ information in combination with web intelligence is sufficient to semantically tag a document, it is important to consider a baseline which takes the full text content of the document for tagging. The full text information is generated from the pdf versions of the test documents. The PDF documents were converted to text files using PDF conversion tools. As a basic pre-processing step, stop-words, non- alphabetical characters and special symbols were removed from the text to generate a bag of word representation of the full text.

For the purpose of comparison, the full text context was used to generate tags using the proposed approach and at the final step, de-noising of tags was done using the proposed algorithm. The purpose of this baseline is to see the effectiveness of the ’short-text’ expansion approach for semantic tagging in comparison to full-text approach.

5.4 Evaluation metrics

Given that the topics/keywords for a document are assigned in natural language, evaluating accuracy of any algorithm for tagging is a challenging task. Though, solutions such as expert’s evaluation exist, but for this project expert assistance was a challenge. In the absence of expert evaluation, we evaluated the results of our approach in the following ways. We evaluated the effectiveness of the proposed approach in the following three ways.

5.4.1 Jaccard similarity with baseline

The Jaccard similarity between two sets A and B is defined as the ratio of the size of the intersection of these sets to the size of the union of the sets. It can be mathematically stated as:

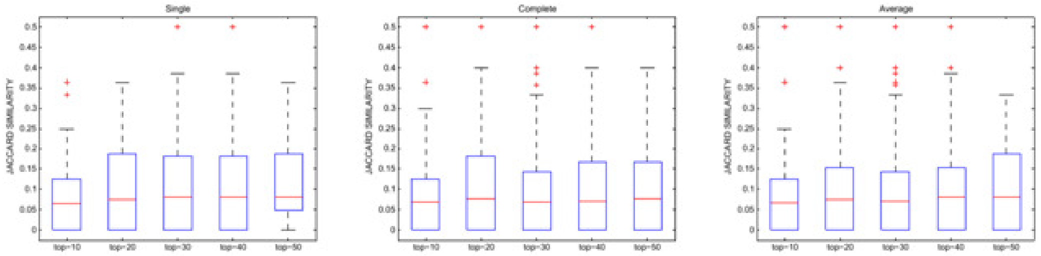


Figure 1. Boxplots showing the distribution of Jaccard Index (Overlap of tags generated from expanded context vs the full content) for 50 documents . The following hierarchical clustering criterion are used : (a) single (b) complete (c) average .

$$J(A,B) = \frac{|A \bigcap B|}{|A \bigcup B|}$$

We used the Jaccard similarity to quantify the similarity between the tags, predicted by the proposed approach versus the baseline approach. This metric represents the magnitude of overlap between the tags generated by the two approaches.

5.4.2 Jaccard index

The Jaccard index is same as the Jaccard similarity except that it is used for a different purpose. Instead of computing the Jaccard index between the results of the proposed approach and the baseline, the Jaccard index for both the approaches is computed with the ground truth tags. This metric gives us the overlap of predicted tags using proposed approach and the baseline with the ground truth tags for each of the test documents. The Jaccard index is averaged over the total number of documents in the test dataset.

5.4.3 Execution time

The final metric for comparing the proposed approach with the baseline is the execution times. Since the main overhead of the approach is in the first step of tag generation due to differences in the sizes of the input context. The execution time is computed as the time taken in seconds to generate tags for the 50 test documents given their input context. For the proposed approach the context is derived using web intelligence whereas for the baseline the context is the full text of the test document. Pre-processing overheads are not taken into account while computing execution timings.

6. RESULTS AND DISCUSSION

This section is sub-categorized into two sections. The first section discusses the quantitative evaluation of the proposed work. In the next section we present a qualitative discussion about the results of the proposed algorithm for some of the documents in the test corpus.

6.1 Quantitative evaluations

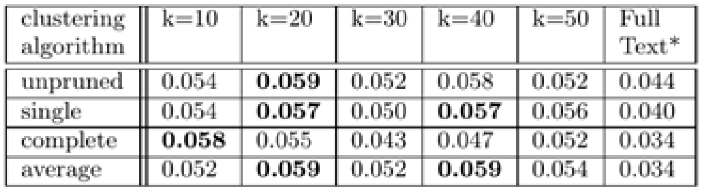
In this section, we describe three experiments conducted to quantitatively evaluate the proposed

approach. The first experiment compares the similarity of the results from the pro- posed and the baseline approaches. The second experiment gives insights about the differences between the proposed approach and the baseline using the ground truth information. The last experiment compares the proposed approach and the baseline based on the execution time performances. These experiments are described as follows.

6.1.1 Experiment 1

Figure 2 shows the distribution of Jaccard similarity for 50 documents for different clustering algorithms. Figure 2 (a), (b) and (c) corresponds to the results of single, complete and average hierarchical clustering based de-noising algorithms. The x-axis of these plots show the variation over k (the top-k headings incorporated in the expanded context). The value of k varies from 10 to 50 in steps of 10. A context made of top-50 web search results are referred as ’top-50’ in the plots. The y-axis shows the Jaccard similarity value. The box in the plots are distribution of the Jaccard similarity values for the 50 test documents. The red bar in the box corresponds to the median of the similarity value, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually as ’+’. As shown in these figures, the value of Jaccard similarity is not very high.
On an average this value is lesser than 0.10 which signifies a low order of similarity between the tags generated using expanded context versus the tags generated using the full text. However, for the sake of comparison, we find that the Jaccard similarity between the expanded context tagging and full text tagging is higher when the value of k is low. There are very few test documents which have a high Jaccard similarity as shown by the outliers. The maximum similarity is 0.5 for almost all the values of k. We also see that using different clustering algorithms do not make a significant difference in the Jaccard similarity.

Table 2: Table showing Jaccard Index measure for the pro-posed approach (varying k in context expansion) and the full content baseline



clustering algorithm	k=10	k=20	k=30	k=40	k=50	Full Text*
unpruned	0.054	**0.059**	0.052	0.058	0.052	0.044
single	0.054	**0.057**	0.050	**0.057**	0.056	0.040
complete	**0.058**	0.055	0.043	0.047	0.052	0.034
average	0.052	**0.059**	0.052	**0.059**	0.054	0.034

6.1.2 Experiment 2

Based on the Experiment 1, we can say that the tags generated using expanded context and the tags generated using the full text do not overlap significantly. However, this experiment does not conclude about the quality of the tags generated by both the approaches. In order to compare the quality of tags generated by both the approaches, we evaluate the results of the proposed approach and the baseline approach using the ground truth tags for the test documents.

The results of this experiment are shown in table 2. As described earlier, we use the Jaccard index to compare between the qualities of the results. The rows in this table correspond to the results obtained by using different clustering algorithms. The first five columns correspond to the expanded context extracted using k as 10, 20, 30, 40 and 50. The last column contains the results for the baseline referred as Full Text since the context consists of the full text of the document. For the first row (unpruned), tags are not de-noised using any algorithm. The Jaccard index of the baseline (Fulltext) with the ground truth is 0.044 whereas the Jaccard index for all the expanded context (proposed approach) over all values of k is greater than 0.50. The highest Jaccard index is 0.059 at k=20.

When we use the single hierarchical clustering algorithm for de-noising, the Jaccard index is only reduced to 0.040 for Fulltext baseline. The Jaccard index in the expanded context with k=20, 40 is 0.057 which is clearly higher to the baseline results. Similarly for the complete hierarchical clustering based de-noising, the Jaccard index is 0.058 for k=10 whereas it is only 0.034 for the full text baseline. The same scenario is found for average hierarchical clustering based de-noising. The Jaccard index is 0.059 for k=20, 40 while it is only 0.034 for Fulltext baseline.

The above described experiment shows a quantitative approach for comparing the quality of resultant tags from the proposed and the baseline approaches. The results shows above, surprisingly, favor the tags generated by the proposed approach which uses only the title/heading information about the document and web intelligence to annotate the document with relevant tags. The baseline approach uses the full text of the document in order to generate annotations. Although the degree of overlap between the predicted tags with the ground truth tags is low (due to the inherent challenge of natural language), the results are useful to show the difference in the quality of predicted tags by the proposed and baseline approaches. An explanation for the observed results can be attributed to the fact that context derived from the web contains a wide spectrum of terms useful for generating generalized tags for the document. While on the other hand, the full text approach uses only the terms local to the specific document which might not be diverse enough to generate generalized tags.

Since an exact match evaluation (as done above) might not fully account for the quality of tags, we demonstrate the results of a few sample documents from the test dataset in the qualitative evaluation section.

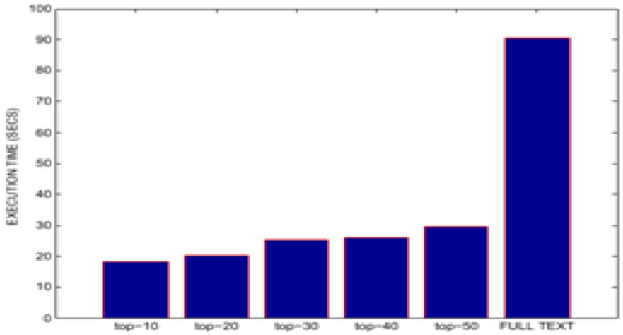


Figure 3: Figure showing execution time comparison for the tag generation step using the expanded context (varying k) vs the full text for 50 documents .

6.1.3 Experiment 3

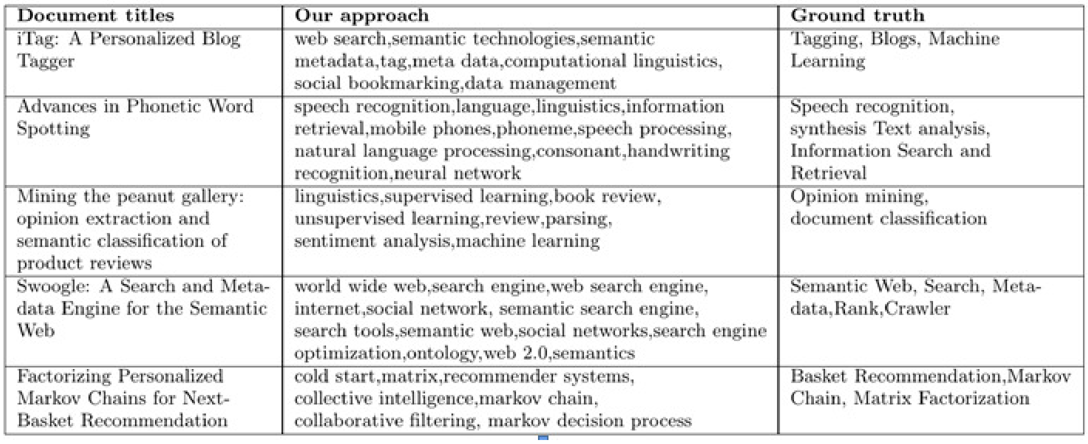
One of the challenges described about using the full text approach for tagging is the issue of time consumption for reading the full text in case the document is large. In Figure 3, we show the results of an experiment conducted to compare the execution time of the tag generation step for the pro- posed and the baseline approaches. The x-axis in the figure shows the expanded context (using different values of k) and the baseline (Full text). The y-axis corresponds to the total execution time (in seconds) for 50 documents. AS shown in the figure, the execution time for the baseline is approximately 90 seconds for 50 documents, whereas the maximum execution time is only 30 seconds for the expanded context where k=50. As shown earlier, that the quality of tags generated using expanded context with k=10 or k=20 is as good as higher values of k. This implies that good quality tags for a document can be generated 4.5 times faster using the proposed approach than using the full text of the document. This shows the effectiveness of the proposed approach to be useful in real time systems.

6.2 Qualitative evaluation

In this section, we discuss the results of the proposed approach by qualitatively analyzing the results of the proposed algorithm. The last section highlighted the performance of the proposed algorithm and quantitatively compared the results with the baseline using the Jaccard index. However, using a quantitative measure like Jaccard fails to account for the subjective accuracy of the tags other than those which do not match the ground truth exactly. Here we analyze the results in a subjective manner.

Table 3 shows the tags, predicted by the proposed approach and the ground truth tags for a few sample documents from the test dataset. The titles shown in this table (in column 1) in general capture the core ingredients of the document. The second column captures the results of the proposed approach and the column 3 captures the ground truth tags.

Table 3: Table showing results for a few of the sample documents. This table shows that several of the topics in the second column (our approach) are very closely related to the keywords in the ground truth (column 3).



Document titles	Our approach	Ground truth
iTag: A Personalized Blog Tagger	web search,semantic technologies,semantic metadata,tag,meta data,computational linguistics, social bookmarking,data management	Tagging, Blogs, Machine Learning
Advances in Phonetic Word Spotting	speech recognition,language,linguistics,information retrieval,mobile phones,phoneme,speech processing, natural language processing,consonant,handwriting recognition,neural network	Speech recognition, synthesis Text analysis, Information Search and Retrieval
Mining the peanut gallery: opinion extraction and semantic classification of product reviews	linguistics,supervised learning,book review, unsupervised learning,review,parsing, sentiment analysis,machine learning	Opinion mining, document classification
Swoogle: A Search and Meta-data Engine for the Semantic Web	world wide web,search engine,web search engine, internet,social network, semantic search engine, search tools,semantic web,social networks,search engine optimization,ontology,web 2.0,semantics	Semantic Web, Search, Meta-data,Rank,Crawler
Factorizing Personalized Markov Chains for Next-Basket Recommendation	cold start,matrix,recommender systems, collective intelligence,markov chain, collaborative filtering, markov decision process	Basket Recommendation,Markov Chain, Matrix Factorization

For the first document in the table (’iTag: A Personalized Bog Tagger’), the keywords (our ground truth) assigned by the used contains terms like ’tagging’, ’blogs’ and ’Machine learning’. The tags generated by the proposed approach are shown in the middle column. Although there are no exact match between the proposed tags and the ground truth tags yet the relevance of the proposed tags is striking. Tags such as ’semantic meta data’, ’social bookmarking’, ’tag’, ’computational linguistics’ are similar others in this list are clearly good tags in this document. Another example is shown in the next row. The ground truth tag ’speech recognition’ exactly match the tag in the proposed list. However, most of the other tags in the list of proposed tags are quite relevant. For example, tags such as ’linguistics’, ’natural language processing’ are closely related to this document. A few tags such as ’mobile phones’, ’consonant, ’hand writing recognition’ may not be directly related. The third example shown in this table also confirms the effectiveness of the proposed approach. The ground truth consists of only two tags: ’opinion mining’ and ’document classification’ while the proposed tag list consists several relevant tags though there is no exact match.

The last two examples shown in this table demonstrate the effectiveness of the approach to expand the annotation with meaningful tags. The fourth example is originally tagged with tags like ’semantic web’, ’search’, ’meta-data’, ’rank’ and ’crawler’. But the proposed tag list consists of highly relevant tags like ’ontology’, ’search optimization’ which capture even the technique used in the particular research document. Similarly, for the last example the not overlapping tags are relevant for annotating the research document.

From the above qualitative analysis, we get a better understanding about the quality of tags generated using the proposed approach. Although it is an interesting challenge to quantitatively describe the quality of the proposed tags, this problem is not addressed in the current version of the work.

7. RELATED WORK

As described earlier, the literature under document annotation can be divided into two broad classes. The first class of approaches studies the problem of annotation using extraction techniques [5, 6]. The main objective of such techniques is to identify important words or phrases from within the content of the document to summarize the document. This class of problem is studied in the literature by several names such as “topic identification” [3],”categorization” [19, 13], “topic finding” [14],”cluster labelling” [18, 16, 21, 24] and as well as “keyword extraction” [5, 6].

Researchers working on these problems have used both supervised and unsupervised machine learning algorithms to extract summary words for documents. Witten et al. [23] and Turney [22] are two key works in the area of supervised key phrase extraction. In the area of unsupervised algorithms for key phrase extraction, Mihalcea and Tarau [17] gave a textRank algorithm which exploits the structure of the text within the document to find key-phrases. Hasan and Ng [8] give an overview of the unsupervised techniques used in the literature.

In the class of key phrase abstraction based approaches. There can be two approaches for document annotation or document classification: single document annotation and multiple document annotation. In the single document summarization, several deep natural language analysis methods are applied. These strategies of document summarization use ontology knowledge based summarization [9, 11]. The ontology sources commonly used are WordNet, UMLS. The second approach widely used in single document summarization is feature appraisal based summarization. In this approach, static and dynamic features are constructed from the given document. Features such as sentence location, named entities, semantic similarity are used for finding documents similarity.

In the case of multi-document strategies, the techniques in- corporate diversity in the summary words by using words from other documents. However, these techniques are limited when the relevant set of documents is not available. Gabrilovich et. al [7] proposed an innovative approach for document categorization which uses of Wikipedia knowledge base to overcome the limitation of generating category terms which are not present in the documents. However, this approach uses the entire content of the document and extend the context using Wikipedia.

8. CONCLUSIONS

In summary, there are three main conclusions in this work. Firstly, we showed an automated approach for tag generation using only a short text information from the document and intelligence from the web. Secondly, we quantitatively evaluated and compared the results of the proposed approach against the baseline approach which uses the full text of the document. We used different metrics to compare and contrast the results. We found that the proposed approach performs better than the baseline approach in terms of the Jaccard index with the ground truth tags. We also found that the proposed approach is at least 3 times faster than the baseline approach and thus, useful for real time system. Thirdly, we evaluated the quality of the proposed tags for documents against the ground truth tags in a qualitative fashion. This analysis reveals the qualitative effectiveness of the proposed approach for meaningful tag generation using only 'short text' information from the document.

There are several areas in this work which we would extend in the near future. One of the areas of improvement in the current work is the de-noising algorithm which uses hierarchical clustering to pruning. However, hierarchical clustering has its limitations and it is worth to explore other algorithms such as density based clustering or some novel anomaly detection algorithm. We would also test the pro- posed approach for other document corpus like news, patents etc. Finally, we also plan to quantitatively validate the accuracy of the results in the case when the results do not exactly match the ground truth.

ACKNOWLEDGEMENTS

This work was supported by the Army Research Laboratory and DARPA and was accomplished under Cooperative Agreement Numbers W911NF-09-2-0053 and W911NF-12-C-0028. We also thank the lab members of DMR lab, University of Minnesota for their valuable suggestions regarding experimental design and comparison.

REFERENCES

[1] AlchemyAPI. Text analysis by alchemyapi, 2013.

[2] R. L. Cilibrasi and P. M. Vitanyi. "The google similarity distance". IEEE Transactions on Knowledge and Data Engineering, Vol 19(3):370–383, 2007.

[3] C. Clifton, R. Cooley, and J. Rennie. "Topcat: data mining for topic identification in a text corpus". IEEE Transactions on Knowledge and Data Engineering, 16(8):949–964, 2004.

[4] O. Corcho. "Ontology based document annotation: trends and open research problems". International Journal of Metadata, Semantics and Ontologies, 1(1):47–57, 2006.

[5] L. Ertöz, M. Steinbach, and V. Kumar. "Finding topics in collections of documents: A shared nearest neighbor approach". Clustering and Information Retrieval, 11:83–103, 2003.

[6] E. Frank, G. W. Paynter, I. H. Witten, C. Gutwin, and C. G. Nevill-Manning. "Domain-specific keyphrase extraction". 1999.

[7] E. Gabrilovich and S. Markovitch. "Overcoming the brittleness bottleneck using wikipedia: Enhancing text categorization with encyclopedic knowledge". In AAAI, volume 6, pages 1301–1306, 2006.

[8] K. S. Hasan and V. Ng. "Conundrums in unsupervised keyphrase extraction: making sense of the state-of-the-art". In Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pages 365–373. Association for Computational Linguistics, 2010.

[9] M. M. Hassan, F. Karray, and M. S. Kamel. "Automatic document topic identification using wikipedia hierarchical ontology". In 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA), 2012, pages 237–242. IEEE, 2012.

[10] Q. He, J. Pei, D. Kifer, P. Mitra, and L. Giles. "Context-aware citation recommendation". In Proceedings of the 19th international conference on World wide web, pages 421–430. ACM, 2010.

[11] S. Jain and J. Pareek. "Automatic topic (s) identification from learning material: An ontological approach". In Second International Conference on Computer Engineering and Applications (ICCEA), 2010, volume 2, pages 358–362. IEEE, 2010.

[12] N. H. Jesse Alpert. "We knew the web was big...", July 2008.

[13] T. Joachims. "Text categorization with support vector machines: Learning with many relevant features". Springer Berlin Heidelberg (pp., 137-142) 1998.

[14] D. Lawrie, W. B. Croft, and A. Rosenberg. "Finding topic words for hierarchical summarization". In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, pages 349–357. ACM, 2001.

[15] M. Ley and P. Reuther. “Maintaining an online bibliographical database: The problem of data quality”. In EGC, pages 5–10, 2006.

[16] C.-Y. Lin. “Knowledge-based automatic topic identification”. In Proceedings of the 33rd annual meeting on Association for Computational Linguistics, pages 308–310. Association for Computational Linguistics, 1995.

[17] R. Mihalcea and P. Tarau. “Textrank: Bringing order into texts”. In Proceedings of EMNLP, volume 4. Barcelona, Spain, 2004.

[18] M. F. Moura and S. O. Rezende. “Choosing a hierarchical cluster labelling method for a specific domain document collection”. New Trends in Artificial Intelligence, pages 812–823, 2007.

[19] F. Sebastiani. “Machine learning in automated text categorization”. ACM computing surveys (CSUR), 34(1):1–47, 2002.

[20] A. Singhal, R. Kasturi, and J. Srivastava. “Automating document annotation using open source knowledge”. In IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013, volume 1, pages 199–204. IEEE, 2013.

[21] S. Tiun, R. Abdullah, and T. E. Kong. “Automatic topic identification using ontology hierarchy”. In Computational Linguistics and Intelligent Text Processing, pages 444–453. Springer, 2001.

[22] P. D. Turney. “Learning algorithms for keyphrase extraction”. Information Retrieval, 2(4):303–336, 2000.

[23] I. H. Witten, G. W. Paynter, E. Frank, C. Gutwin, and C. G. Nevill-Manning. “Kea: Practical automatic keyphrase extraction”. In Proceedings of the fourth ACM conference on Digital libraries, pages 254–255. ACM, 1999.

[24] O. Zamir and O. Etzioni. “Web document clustering: A feasibility demonstration”. In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, pages 46–54. ACM, 1998.

AUTHORS

Ayush Singhal was born in India in the year 1990. He is currently a second year PhD student in the computer science department at the University of Minnesota, USA. He completed his under graduation from the Indian Institute of Technology Roorkee, India in 2011. His major is computer science. His current research interests are data mining, information retrieval, web mining and social network analysis.As an under graduate he has co-authored two conference papers in prestigious national and international IEEE conference. He has also published a journal article in Springer journal (Real time image processing). He has been working as a research assistant in the University of Minnesota for 2 years now. He has also worked in IBM Research labs, New Delhi India as a summer intern in year 2010.

Jaideep Srivastava received the Btech degree in computer science from The Indian Institute of Technology, Kanpur, India, in 1983, the MS and PhD degrees in computer science from the University of California, Berkley, in 1985 and 1988 respectively. He has been on the faculty of the Department of Computer Science and Engineering of the University of Minnesota, Minneapolis, since 1988 and is currently a professor.He served as a research engineer with Uptron Digital Systems in Lucknow, India, in 1983. He as published more than 250 papers in refereed journals and conference proceedings in the areas of databases, parallel processing, artificial intellig3ence, multimedia and social network analysis; and he has delivered a number of invited presentations and participate in panel discussions on these topics. His professional activities have included service on various program committees and he has refereed papers for varied journals and proceedings, for events sponsored by the US National Science Foundation. He is a Fellow of the IEEE, and a Distinguished Fellow of Allina Hospitals’ center for Healthcare Innovation. He has given over 150 invited talks in over 30 countries, including more than a dozen keynote addresses at major conferences.