Annotating Text Segments Using a Web-Based Categorization Approach

Hsin-Chen Chiao[1], Hsiao-Tieh Pu[2], and Lee-Feng Chien[1]

[1] Institute of Information Science, Academia Sinica, Taipei, Taiwan 115
{hcchiao, lfchien}@iis.sinica.edu.tw
[2] Graduate Institute of Library & Information Studies, National Taiwan Normal University, Taipei, Taiwan 106
htpu@ntnu.edu.tw

Abstract. Conventional automatic text annotation tools mostly extract named entities from texts and annotate them with information about persons, locations, and dates, etc. Such kind of entity type information, however, is insufficient for machines to understand the context or facts contained in the texts. This paper presents a general text categorization approach to categorize text segments into broader subject categories, such as categorizing a text string into a category of paper title in Mathematics or a category of conference name in Computer Science. Experimental results confirm its wide applicability to various digital library applications.

1 Introduction

Text mining can be used to add value to unstructured data, like documents in digital library collections. [1] In general, current tools for automatic text annotation mostly extract named entities from texts and annotate them with information about persons, locations, dates and so on. [2] However, this kind of entity type information is often insufficient for machines to understand the facts contained in the texts, thus preventing them from implementing more advanced or intelligent applications, such as text mining. [3] In this paper, we try to remedy this problem by presenting a more generalized text categorization approach which is pursued to categorize text segments, i.e., meaningful word strings including named entities and other important text patterns such as paper titles and conference names, into broader subject categories.

Named Entity Recognition (NER) is an important technique used in many intelligent applications such as information extraction, question answering and text mining. The NER task consists in identifying phrases in text, which are often short in length, e.g., single words or word bigrams, into certain types such as organizations, persons and locations. To deal with such task, effective techniques are required to delimit phrases and exploit various evidences of the candidate strings to classify entities. Fewer investigations or research were found to recognize longer strings, such as paper titles and conference names, and classifying them into broader subject categories of concern. Examples include categorizing a paper title into Mathematics category or a conference name into Computer Science category. This paper addresses the problem

E.A. Fox et al. (Eds.): ICADL 2005, LNCS 3815, pp. 323 – 331, 2005.

of text segment categorization and presents a feasible approach using the Web as an additional knowledge source.

In this paper, a text segment is defined as a meaningful word string that is often short in length and represents specific concept in a certain subject domain, such as a keyword in a document set and a natural language query from a user. Text segments are of many types, including word, phrase, named entity, natural language query, news event, product name, paper or book title, etc. Categorizing short text segments is a difficult problem given that, unlike long documents, short text segments typically don't contain enough information to extract reliable features. For longer documents, their subject information can be represented based on the composed words and the similarities to a classifier can be estimated based on the common composed words. However, for text segments, their subject information cannot be simply judged by using the same way due to the fact that text segments are usually short and don't contain enough information in the composed words. Thus, the most challenging task is to acquire proper features to characterize the text segments. For those text segments extracted from documents, e.g., key terms from documents, the source documents can be used to characterize the text segments. However, in real-world cases, such as in dealing with search engine query strings, there may not exist sufficient relevant documents to represent the target segments. In other words, the lack of domain-specific corpora to describe text segments is usually the case in reality.

Fortunately, the Web, as the largest and most accessible data repository in the world, provides rich resources to supplement the insufficiency of information suffered by various text segments. Many search engines constantly crawls Web resources and provides relevant Web pages for large amounts of free text queries consisted of single terms and longer word strings. The major idea of the proposed approach is to use the Web search result snippets to extract related contextual information as the source of features for text segments. In other words, the proposed approach incorporates the search result snippets returned from search engines into the process of acquiring features for text segments. Often there are some text segments too specific to obtain adequate search results using current keyword-matching-based search engines. This motivates our exploration of a better query processing technique, named query relaxation, which is designed to acquire more relevant feature information for long text segments through a bootstrapping process of search requests to search engines. Initial experiments on categorizing paper titles into Yahoo!'s Computer Science hierarchy has been conduced and the experimental results show the potential and wide adaptability of the proposed approach to various applications.

2 Related Work

2.1 Named Entity Recognition

Effective techniques are required in NER to delimit phrases and exploit various evidences of the candidate strings to classify entities. It has been a well-accepted principle that two different types of evidences, i.e., internal and external evidences, are keys in clarifying the ambiguities and improving the robustness and portability. For example, the internal evidences, such as capitalization, are features found within the

candidate string itself; while the external evidences, such as neighboring words associations, are derived by gathering the local context into which the string appears. A number of approaches have been developed for utilizing external evidences to find functional-similar words and identifying named entities. Usually these approaches rely on analysis of the considered objects' contextual information obtained from tagged corpus. [4] Instead of using tagged corpus for categorizing word- or phrasal-level objects, the proposed approach exploits Web resources as a feature source to categorize text segments, which might be longer in length, into broader subject categories. Our research assumes that the text segments are formed with a simple syntactic structure containing some domains-specific or unknown words. Either conventional syntactic sentence analysis or complete grammatical sentence analysis may not be appropriate to this case.

2.2 Text Categorization

Text categorization techniques are often used to analyze relationships among documents. [5] However, as previously mentioned, there is a great difference between document categorization and text segment categorization. Documents normally contain more information than text segments do. The similarity between a document and a target category can be estimated based on the difference in the distribution of the words contained in document itself and the training set of the category; whereas the similarity between a short text segment and a target category cannot be estimated in this way. Further, conventional text categorization techniques assume manually-labeled corpora are ready and can be used for training process. In reality, labeling the corpus is laborious and may suffer from the problem of subjectivity. Using the Web as a corpus source proves a better alternative. Our previous work has proposed an approach to train classifiers through Web corpora to build user-defined topic hierarchies. [6] The proposed approach in this paper extends the previous work, and focuses on the text segment categorization problem.

2.3 Text Mining and Web Corpora

Our research is also related to the work concerning with the knowledge discovery in huge amounts of unstructured textual data from the Web. [7] To name a few related research here, such as automatic extraction of terms or phrases, [8] the discovery of rules for the extraction of specific information patterns, [9] and ontology construction based on semi-structured data. [10] Different from previous works, the proposed approach is to categorize text segments via mining search result pages.

3 The Approach

This section first defines the problem, and then introduces the proposed approach. Given a set of subject categories, $C = \{c_1, c_2, \ldots, c_n\}$, a collection of text segments $T = \{t_1, t_2, \ldots, t_m\}$, and also a mapping $M : T \rightarrow C$ that describes the correct category a text segment is supposed to be assigned with. The major concern is to design a one-to-one mapping scheme M': $T \rightarrow C$ that the maximal size of the correct result set is $CRS = \{t_i \mid t_i \text{ in } T, M'(t_i) = M(t_i)\}$. The approach is essentially composed of two

computational modules: feature extraction and text segment categorization. The approach exploits highly ranked search result snippets retrieved from search engines as the feature sources. The feature extraction module collects features for the text segment of concern. The text segment categorization module decides appropriate categories for the text segment. Detailed discussion of each module is presented in the following subsections.

3.1 Feature Extraction and Representation

To decide the similarity between a text segment and a target subject category, a representation model is necessary to describe it characteristics. As previously mentioned, a text segment cannot offer sufficient feature terms by itself. In other words, calculating the distance between the text segment and a target category directly is not possible. To overcome this problem, the approach sends the text segment as a query to search engines and use the returned pages as its feature source. Note that, instead of the whole page, only the snippets were used as the sources to save a large number of page accesses. The approach adopts the vector space model to describe the features of both text segments and thematic categories. Suppose that, for each query q (in fact a text segment or some Boolean expressions of category names), the approach collects up to N_{max} search result snippets, denoted as SRS_q. Each query can be then converted into a bag of feature terms by applying normal text processing techniques, e.g., removing stop words and increasing stemming, to the contents of SRS_q. Let T be the feature term vocabulary, and t_i be the *i-th* term in T. With a simple processing, a query q can be represented as a term vector v_q in a $|T|$-dimensional space, where v_{qi} is the weight t_i in v_q. The term weights in this work were determined according to one of the conventional *tf-idf* term weighting schemes, in which each term weight v_{qi} is defined as:

$$v_{q,i} = (1 + \log_2 f_{q,i}) \text{ x } \log_2(n / n_i),$$

where f_{qi} is the frequency of t_i occurring in v_q's corresponding feature term bag, n is the total number of category objects, and n_i is the number of category objects that contain t_i in their corresponding bags of feature terms. The similarity between a text segment and a category object is computed as the cosine of the angle between the corresponding vectors, i.e.,

$$sim(v_a, v_b) = cos(v_a, v_{b)},$$

3.2 Text Segment Categorization

Given a new text segment t, the approach determines a set of categories C_t that are considered as t's most related categories. As discussed in previous section, the candidate text segment t is represented as a feature vector v_t. For this categorization task, a kNN approach was used. kNN has been found to be an effective classification approach to a broad range of pattern recognition and text classification problems. Using the kNN approach, a relevance score between t and candidate category object C_i is determined by the following formula:

$$r_{kNN}(t, C_i) = \sum_{v_j \in R_k(t) \cap C_i} sim(v_t, v_j)$$

where $R_k(t)$ represents t's k most-similar category objects, measured by a *sim* function, in the whole collection. The categories a text segment being assigned with are determined by either a predefined number of most-relevant clusters or a threshold used to pick those clusters having scores higher than that of the specified threshold value. The performance evaluation of the proposed approach was mainly based on the extraction of five most-relevant categories as candidates.

3.3 Query Relaxation

Sometimes there exist text segments that are too specific to obtain adequate search results using current keyword-matching-based search engines. Insufficient snippets may cause the obtained information sparse and not so unreliable, and may even decrease the relevance measurement among text segments. The case of retrieving inadequate search results mostly occurs when dealing with long text segments, e.g., paper titles and natural language queries. Compared with a short text segment, a long segment contains more information, i.e., with more terms, and it's rather difficult to obtain documents exactly matching all of the terms. However, as a long text segment contains more information, not all terms in the segment are equally informative to its intended topic(s). This motivates our invention of a query processing technique, named query relaxation, to acquire more relevant feature information for long text segments through a bootstrapping process of search requests to search engines.

q=Polynomial-Time Reinforcement Learning of Near-Optimal Policies
  q^1=Reinforcement Learning Near-Optimal Policies
  q^2=Reinforcement Learning Policies
q=Named Entity Recognition using an HMM-based Chunck Tagger
  q^1=Named Entity Recognition HMM-based Tagger
  q^2=Named Entity Recognition Tagger
q=A digital library of conversational expressions: helping profoundly disabled users communicate
  q^1=digital library conversational expressions helping disabled users communicate
  q^2=digital library conversational expressions helping users communicate

Fig. 1. Examples of paper title and their relaxed versions

To clarify the idea of query relaxation, let's take the title of this paper as an example of long text segment: "Annotating Text Segments Using A Web-based Categorization Approach." Suppose that one needs to select a subset of terms as the query that can mostly represent the topical concept of this segment, one most probably selects those of "Annotating Text Segments." If one needs to further reduce the sub-segment "Annotating Text Segments," "Annotating Text" seems a better choice. Though this selection process may not be always feasible and may depend on the decisions made, the idea shows that the textual part of a long text segment can be effectively reduced or relaxed. The reduced segment represents a concept that is still close to (or usually it is broader than) the main topical concept of the original segment, and it usually can retrieve more search results due to the reduced segment holds fewer terms. The above

example suggests a possible approach in an inclusion manner, i.e., to select a subset of terms that are most informative from the given text segment. Instead of following such inclusion manner, our approach was designed in an exclusion manner. In other words, when the search results of the given text segment are not adequate, a single term is removed from the segment, and the rest form a new query to search engines. The newly retrieved search results are then augmented into the set of the original search results. For those overlapping entries, they will be deleted and not added. This relaxation process is repeated until the obtained information is considered enough. For illustration, figure 1 shows several examples of paper titles with their relaxed versions obtained using the proposed query relaxation technique.

4 Experiments

To assess the performance of the proposed approach, some initial experiments have been conducted. We used the Yahoo! Computer Science hierarchy as the subject categories of concern. In the hierarchy, there are totally 36 second-level, 177 third-level, and 278 fourth-level categories, all rooted at the category of "Computer Science". A data set consisting of the academic paper titles were collected from six computer science conferences held in 2002. The experiment tried to categorize them into the 36 first-level categories, such as "Artificial Intelligence" and "Operating Systems". Table 1 lists the relevant information of this paper data set. For each category, we created a text classifier using a training corpus obtained via taking the category name itself and each of its subcategory names as a query to retrieve search result snippets respectively, which is a process similar to that of extracting feature sources for text segments.

To evaluate the categorization accuracy, each conference was assigned to the Yahoo! categories to which the conference was considered to belong, e.g., AAAI'02 was assigned to "Artificial Intelligence", and all the papers from that conference were unconditionally assigned to that category. Notice that this might not be absolutely correct categorization strategy as some papers in a conference may be more related to other domains than the ones assigned. To make the experiment easier to implement, we made this straightforward assumption.

Tables 2 shows the results of the achieved top 1-5 inclusion rates, where the top *n* inclusion rate is the rate of the test text segments (paper titles) whose highly ranked *n*

Table 1. The information of the paper data set

Conference	# Papers	Assigned Category
AAAI'02	29	CS: Artificial Intelligence
ACL'02	65	CS: Linguistics
ICML'02	87	CS: Artificial Intelligence
JCL'02	69	CS: Lib. & Info. Sci.
SIGCOMM'02	25	CS: Networks
SIGGRAPH'02	67	CS: Graphics

candidates contain the correct category. This experiment was conducted without using the query relaxation technique. From Table 2, it shows that the achieved accuracy was promising.

Table 2. Top 1-5 inclusion rates for categorizing paper titles

Conference	Top-1	Top-2	Top-3	Top-4	Top-5
AAAI'02	.6897	.7586	.8621	.8966	.9301
ACL'02	.5321	.7077	.7692	.8	.8153
ICML'02	.5172	.6437	.7701	.8161	.8391
JCDL'02	.2753	.4493	.4927	.5072	.5217
SIGCOMM'02	.88	1.0	1.0	1.0	1.0
SIGGRAPH'02	.8599	.9552	.9552	.9701	.9701
AVG	.5965	.7193	.7690	.7953	.8187

Table 3 further lists some wrongly categorized examples and it can be observed that not all of the miss-categorized papers might be incorrect. In some cases, they were more related to the result subject categories than those we assigned to them. This experiment reveals a great potential of using the proposed approach to categorizing paper titles and organizing academic papers on the Web.

Table 3. Selected examples of miss-categorized paper titles.

Paper Title	Conference	Target Cat.	Top-1	2	3	4	5
A New Algorithm for Optimal Bin Packing	AAAI	AI	ALG	AI	MOD	COLT	DNA
(Im)possibility of Safe Exchange Mechanism Design	AAAI	AI	NET	SC	LG	DB	MD
Performance Issues and Error Analysis in an Open-Domain Question Answering System	ACL	LG	AI	LG	ALG	DC	SC
Active Learning for Statistical Natural Language Parsing	ACL	LG	AI	LG	NN	COLT	ALG
Improving Machine Learning Approaches to Coreference Resolution	ACL	LG	AI	LG	ALG	FM	NN
A Language Modelling Approach to Relevance Profiling for Document Browsing	JCDL	LIS	AI	UI	LG	LIS	ALG
Structuring Keyword-based Queries for Web Databases	JCDL	LIS	AI	LIS	DB	ALG	ARC
A Multilingual, Multimodal Digital Video Library System	JCDL	LIS	LG	UI	LIS	ECAD	NET
SOS: Secure Overlay Services	SIGCOMM	NET	SC	NET	MC	OS	DC

Abbreviation List:

AI :Artificial Intelligence	DNA :DNA-Based Computing	MOD:Modeling
ALG :Algorithms	ECAD:Electronic Computer Aided Design	NET :Networks
ARC :Architecture	FM :Formal Methods	NN :Neural Network
COLT:Computational Learning Theory	LG :Linguistics	OS :Operating Systems
DB :Databases	LIS :Library and Information Science	SC :Security
DC :Distributed Computing	MC :Mobile Computing	UI :User Interface

It was also observed in the experiment that many paper titles were too long and the search engine did not provide a sufficient number of Web pages, which undoubtedly lowered the accuracy rate. Our study was thus interested in whether the query relaxation technique could overcome this problem. We adjusted the values of N_{min} and N_{max}, i.e., the minimal and maximal number of Web pages to describe the text segments, and conducted the same experiment again. Table 4 lists the archived result. Note that the result achieved without using the query relaxation technique can be taken as applying query relaxation vacuously, i.e., $N_{max} = 100$, $N_{min} = 0$. From this table, it can be observed the query relaxation technique did help to boost the accuracy rate, though

the extent of the improvement is limited. Another interesting observation can be made from the table is that the size of the training data to describe text segments doest not necessarily bear a positive influence on the performance of our approach. Fewer pages sometimes describe a text segment more precisely – a possible conclusion considering that the more pages, the more noises may occur.

Table 4. Top 1 inclusion rate applying the proposed query relaxation technique

#Snippets	N_{min} = 0	= 25	= 50	= 75
Nmax = 25	.6550	N/A	N/A	N/A
50	.6199	.6374	N/A	N/A
75	.6082	.6082	.6140	N/A
100	.5965	.5965	.5965	.6082

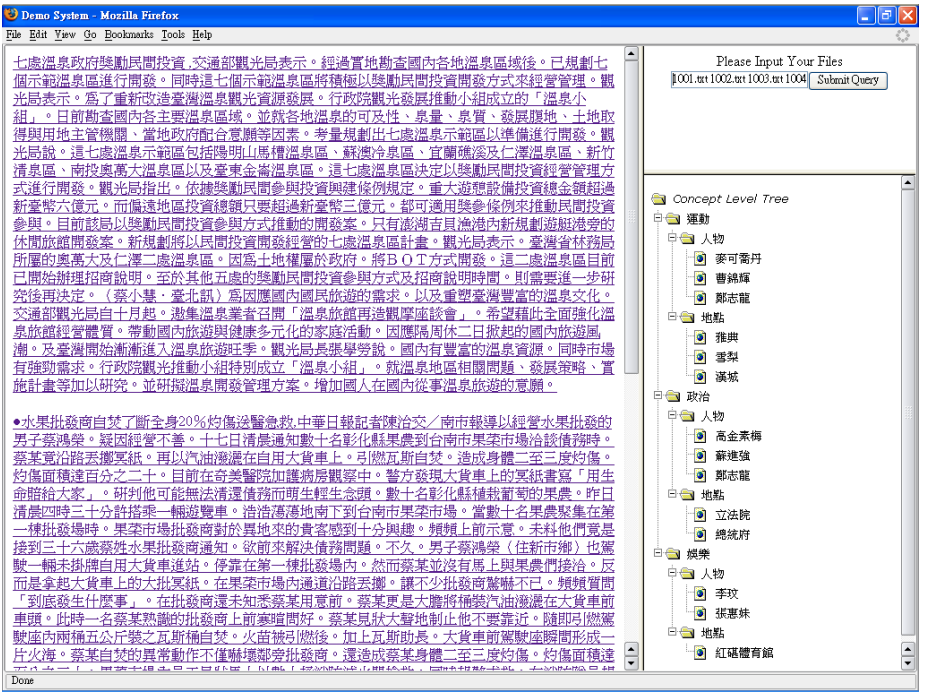

Fig. 2. A prototype information summarization system allowing users to browse documents with the categorized key terms and user-defined categories

The proposed approach is independent of language differences. Currently, the approach has been applied to developing a system called LiveSum, which allows users to browse Chinese documents through the categorized key terms and user-defined categories. As shown in Figure 2, the system extracts key terms from documents (a set of documents collected in a digital library or retrieved from a search engine), and classifies them into user-defined categories. The corresponding classifiers can be

developed using the Web mining approach proposed by Huang, Chuang & Chien (2004). As noted, the system accepts a list of file names as given at the upper-right corner. These documents are displayed at the left part of the browser, and some key terms are extracted and classified into user-defined categories as at the lower-right corner. This provides a new way of information summarization. As the authors observed, it can benefit a lot when a user wants to quickly browse the important concepts embedded in a set of documents.

5 Conclusion

In this paper, we have addressed the problem of text segment categorization and presented a feasible approach dealing with the problem by using the Web as an additional knowledge source. The proposed approach is able to categorize text segments into broader subject categories and is more generalized than conventional named entity recognition approaches. Some initial experiments on categorizing paper titles into Yahoo!'s Computer Science hierarchy has been conducted and the achieved experimental results confirm the potential and its wide adaptability to various digital library applications.

References

1. Witten, I.H., et al.: Text Mining in a Digital Library. *International Journal on Digital Libraries* 4:1 (2004) 56-59.
2. Zhou, G.D., Su, J.: Named Entity Recognition Using an HMM-based Chunk Tagger. *Proceedings of the 40th Annual Meeting of the ACL* (2000) 473-480.
3. Hearst, M.: Untangling Text Data Mining. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics* (1999).
4. Banko, M., Brill, E.: Scaling to Very Large Corpora for Natural Language Disambiguation. *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics* (2001) 26-33.
5. Cohen, W., Singer, Y.: Context-sensitive Learning Methods for Text Categorization. *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (2001) 307-315.
6. Huang, C.C., Chuang, S.L., Chien, L.F.: LiveClassifier: Creating Hierarchical Text Classifiers through Web Corpora. *Proceedings of the 2004 World Wide Web Conference (WWW'04)* (2004).
7. Kosala, R., Blockeel, H.: Web Mining Research: A Survey. *ACM SIGKDD Explorations*, 2:1 (2000) 1-15.
8. Feldman, R., et al.: Maximal Association Rules: A New Tool for Mining for Keyword Co-occurrences in Document Collections. *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining* (1997) 167-170.
9. Soderland, S.: Learning Text Analysis Rules for Domain-specific Natural Language Processing. Ph.D. thesis, technical report UM-CS-1996-087 University of Massachusetts, Amherst (1997).
10. Agirre, E., Ansa, O., Hovy, E., Martinez, D.: Enriching Very Large Ontology Using the WWW. *Proceedings of ECAI 2000 Workshop on Ontology Learning* (2000).