Automatically Acquiring Training Sets for Web Information Gathering

Xiaohui Tao, Yuefeng Li, Ning Zhong*, Richi Nayak
Faculty of Information Technology, Queensland University of Technology, Australia
**Department of Systems and Information Engineering, Maebashi Institute of Technology, Japan*
{*x.tao, y2.li, r.nayak*}*@qut.edu.au, *zhong@maebashi-it.ac.jp*

Abstract

The traditional techniques rely on human effort to acquire training sets, which is expensive and inefficient. In this paper we present an alternative method to automatically acquire training sets without heavy investment of user efforts. The proposed method tends to fill a gap for effectiveness of using Web data in Web mining, and contributes to Web information gathering. The evaluation shows that the method is adequate to yield an promising achievement.

1 Introduction

Over the last decade, we have witnessed an explosive growth in the data available on the Web. However, there are two fundamental issues regarding the effectiveness of using Web data: information mismatch and information overload. These issues lead to a challenge in artificial intelligence (AI) community of "what information gathering has to do with AI" [4]. Web Intelligence (WI) is a new direction which may possibly lead to the solution of these issues [6] [13]. Currently, there are three main directions in WI: Web mining, adaptive Web systems, and information foraging agents. Based on the type of Web data analysed, Web mining is sub-classified in Web usage mining, Web structure mining, Web user profile mining, and Web content mining [6] [9]. In Web profile mining, there are two different diagrams: data diagram and information diagram. Data diagram is the discovery of interesting registration data and customer profile portfolios. Information diagram is the discovery of interesting topics of Web user information needs [7]. These topics may be described by a set of Web documents called training set in Web user profile mining.

In order to acquire a training set, the traditional techniques require users to invest a great deal of efforts [7]. Users are requested to read a set of documents, and provide feedbacks of either positive or negative of the documents to a given topic. Such training sets justified by users manually are called "perfect" training sets in this paper. However, Web users may not like to invest great efforts while performing a search [3]. Sometimes a document covers multiple topics, which may cause a mono-judgement of "YES" or "NO" difficult to make by a user.

In this paper, we propose an alternative method to solve these problems and acquire training sets automatically without users' heavy involvement. The method is developed based on the observations of the existing Web, and employing the Web regularities to support the judgments of positive or negative to the training documents gathered from the Web. The works aim to create a bridge between Web mining and the effectiveness of using Web data, and contribute to Web information gathering. The proposed method is evaluated by the experiments performed on Reuters Corpus Volume 1 (RCV1)[1] data set. The results indicate that the proposed method has made great achievement. The remainder of the paper is structured as the follows. Section 2 present the definitions and the proposed automatically acquiring training sets method. Section 3 discusses the evaluation, and Section 4 presents the related works and makes the conclusions.

2 Automatically Acquiring Training Sets

While acquiring a training set, users can do a great job manually. A user reads each document, and makes a binary judgment of "YES" or "NO" indicating positive and negative of the document. However, in terms of machine learning, a computer can not make the same decision effectively because it does not have any expertise as what a user holds. In terms of human, such expertise is learned from the daily life. Whereas in terms of Web information gathering, we argument that such expertise may be learned from the observations of the existing Web and the related Web regularities, such as: 1. An information need may cover multiple concepts, and the coverage on different concepts may be differ-

[1] Reuters Corpus, http://about.reuters.com/researchandstandards/corpus/.

ent. Sometimes these concepts may overlap. 2. Different Web search engines hold different performances; 3. Different documents indexed on the list returned from a Web search engine holds different certainty degrees to a given query. Based on the observations we may be able to automatically acquire some Web documents and measure their importance to an information need. We propose a method that analyses an information need and identifies its related subjects, then uses the subjects to gather a set of Web documents using a selected Web search engine. The method measures the certainties of the documents, and then makes a float type positive (or negative) judgment to the document based on it. The follows will discuss the method step by step.

2.1 Definitions

Let $D = \{d_1, d_2, \dots, d_m\}$ be a set of documents, $D[d]$ be an index position of d in D. Let $Q = \{q_1, q_2, \dots, q_f\}$ be a set of queries, and $q = \{t_1, t_2, \dots, t_g\}$ be a set of terms. A set of terms is referred to as a $termset$. Therefore, a q can be represented as $termset(q)$, same as a document $termset(d)$.

Let S be a hypothesis space, $s \in S$ is a subject in S. $sem(s)$ is a specific semantic space referred by a subject s and can be represented by $termset(s)$. We have $sem(\{s_i\}) = sem(s_i)$ and $sem(s_i) \subseteq sem(S)$. Therefore, let S_1 be a set of subjects and $S_1 \subseteq S$, S_2 be another set of subjects and $S_2 \subseteq S$, we may have:

$$\begin{aligned} sem(S_1 \wedge S_2) &= sem(S_1) \cap sem(S_2); &(1)\\ sem(S_1 \vee S_2) &= sem(S_1) \cup sem(S_2). &(2)\end{aligned}$$

2.2 Topic Signature Identification

Signature yields a means of the semantic space referred by a user's information needs. In order to appropriately interpret a Web user's information needs, it is necessary to identify the signature of a topic. A signature consists of two types of subjects: *feature subjects* and *noise subjects*. Feature subjects are referred to a semantic space that a specified topic can be best described and discriminated from others; whereas noise subjects are referred to a semantic space that is of little help to distinguish, or sometimes even makes confusion to a topic. Feature subjects determine the search terms for positive training set acquiring. Noise subjects determine the search terms for negative training set acquiring. Subjects may overlap each other. Such overlaps, especially the overlaps highlighted by both of the feature subjects and noise subjects need to be identified, as these overlaps are the confusing semantic space to a topic. Thereafter, the signature sig of a topic is identified as:

$$sig(topic) = sem(S^f_{topic}) - sem(S^f_{topic} \wedge S^n_{topic}) \quad (3)$$

where S^f_{topic} is a set of feature subjects and S^n_{topic} is a set of noise subjects about $topic$ respectively.

Feature subjects and noise subjects can be identified by the descriptions and the narratives provided by a user. It may be desirable to ask users to provide such small piece of information without a great deal of involvement. Because the conceptual coverage of each subject referred by a topic is varying, a subject's belief (or disbelief) to the topic varies as well. Thus, feature subjects and noise subjects should be identified with a certainty factor value based on how strong a subject is for or against a topic (The certainty factor model is borrowed from the MYCIN expert system developed by Stanford, readers may see [2] for details). A certainty factor $CF(topic|s)$ is determined by the belief $MB(topic|s)$ (how strong s is for $topic$) and the disbelief $MD(topic|s)$ (how strong s is against $topic$) of a subject s to a topic:

$$CF(topic|s) = MB(topic|s) - MD(topic|s). \quad (4)$$

As a certainty factor value increases toward 1, a s becomes more strong of supporting the $topic$, and therefore can be classified as a feature subject. Whereas a certainty factor value is lower than (including) 0 and decreases toward -1, a s becomes more strong of being against the $topic$, and can be classified as a noise subject. Based on the certainty factors, a $topic$ may then be interpreted as a set of patterns $(s, |CF(topic|s)|)$, where

$$\begin{cases} s \in S^f_{topic} & \text{if } CF(topic|s) > 0; \\ s \in S^n_{topic} & \text{if } CF(topic|s) \leqslant 0. \end{cases} \quad (5)$$

Each s produces a query q to either the positive query set Q^+ or negative query set Q^- by $termset(q) = termset(s)$, depending on $s \in S^f_{topic}$ or $s \in S^n_{topic}$, respectively.

2.3 Web Search Engine Selection

In order to gathering the best Web data, some factors of an appropriate Web search engine may be specified as follows. If an incoming information need is for general topics, ideally the selected Web search engine may cover Web data in multiple domains. If the topic of information needs is identified focusing on a particular domain, then the Web search engines focusing on that particular domain may be selected. Ideally the selected Web search engine may be recognized by common Web users, so that the acquired training sets are biased towards common knowledge. Moreover, the performance of the selected Web search engine should be with high precision, so that the acquired training sets are of high relevance to a topic. The precision achieved by a Web search engine can be measured by sending a training query to it to search and then analysing the returned results. A common measure of the precision $\wp$ for a search engine θ is as $\wp(\theta, \kappa) = \frac{|D^{'}|}{\kappa}$, where $|D^{'}|$ is the number of counted relevant documents before reaching cutoff κ, and $|D^{'}| \leqslant \kappa$.

2.4 Training Set Gathering

A certainty degree of a document to a topic is firstly affected by its index. The search results provided by a Web search engine are indexed. The documents indexed on the top of list are more relevant to a given topic than the documents indexed on the bottom. A certainty degree is also affected by the precision $\wp$ of the selected Web search engine. If d is in the range of cutoff κ_1 and $\wp(\theta, \kappa_1) = 0.9$, we may say its relevance evidence is stronger than a document in the range of cutoff κ_2 and $\wp(\theta, \kappa_2) = 0.8$. Based on these, with Equation (3) and (4), a certainty degree CD of document d to a *topic* is determined by the sum of belief $MB(s|d)$ measuring how d is supporting feature subjects and the sum of disbelief $MD(s|d)$ measuring how d is supporting noise subjects:

$$\begin{aligned} CD(topic|d) = {} & \textstyle\sum_{s \in S^f_{topic}} MB(s|d) \\ & - \textstyle\sum_{s \in S^n_{topic}} MD(s|d); \end{aligned} \qquad (6)$$

where

$$MB(s|d) = \begin{cases} \gamma & \text{if } d \in D^+ \\ 0 & \text{otherwise;} \end{cases} \qquad (7)$$

$$\begin{aligned} \gamma = {} & |CF(topic|s)| \times \wp(\theta, \kappa) \\ & \times (\tfrac{k-(D^+[d] mod(k))+1}{k}); \end{aligned} \qquad (8)$$

where we call γ a belief value assigned to d, which is determined by the factors of the certainty factor of s, the precision of Web search engine θ, and the document's index $D^+[d]$. D^+ is a positive candidate set, which is gathered by θ using $q \in Q^+$, where $termset(q) = termset(s)$. k is a static number of how many documents in each cutoff. By using the same equations as Equation (7) and (8), $MD(s|d)$ is calculated, where the γ is a disbelief value, and D^+ is changed to D^- for a negative candidate set.

A final training set D^τ acquired consists of a positive set $D^{\tau+}$ and a negative set $D^{\tau-}$. The element in the training set is pattern $(d, |\aleph|)$ where $\aleph = CD(topic|d)$, and

$$\begin{cases} D^{\tau+} = D^{\tau+} \cup \{(d, |\aleph|)\} & \text{if } \aleph > 0 \\ D^{\tau-} = D^{\tau-} \cup \{(d, |\aleph|)\} & \text{otherwise.} \end{cases} \qquad (9)$$

The higher $CD(topic|d)$ is, more confident d is to support *topic*. Whereas once less than 0, the lower $CD(topic|d)$ is, more confident d is against *topic*.

By applying the proposed method users no longer need to manually read a set of supplied documents and provide their binary judgments. No heavy involvement of user effort is necessary for acquiring training sets.

3 Evaluation

By using the proposed method, since the judgments provided for the documents in the training set are no longer traditional binary type but a float digit indicating the certainty degree supporting or against a information need, the traditional information retrieval models need to be modified in order to apply the proposed model. Such modification may be presented by using a sample of traditional *probabilistic* model, which is a popular method for supervised estimating of term weights. The original term weight in *probabilistic* model is estimated by how often a term appears or not appears in positive documents and negative documents respectively [5]. By applying the proposed training sets automatic acquiring method, the frequency r of the positive documents containing a term t is refined as:

$$r = \sum_{d \in D^\triangle} |CD(topic|d)| \qquad (10)$$

where $D^\triangle = \{d | d \in D^{\tau+}, t \in d\}$. And the frequency R of the positive documents in the training set may be refined using the same equation as (10), where the $D^\triangle$ is changed to $D^{\tau+}$ for all the documents in the positive training set.

In the evaluation, the Reuters Corpus Volume 1 (RCV1) is chosen as the testbed, and the TREC-11[2] 2002 Filtering track is chosen as the evaluation scheme. RCV1 is an archive of 806,791 documents produced by Reuters journalists between 1996-08-20 and 1997-08-19. And RCV1 is the testbed used in the TREC-11 2002 Filtering track as well. The TREC-11 Filtering track aims to evaluate the methods of persistent user profiles building and documents classification. Since the proposed Web information gathering method fells into this track, we believe the evaluation scheme of TREC-11 Filtering track along with the RCV1 testbed may be the best design for our experiments.

The experiment design is described as the follows. The TREC-11 Filtering track has 100 topics, 50 of them are constructed manually by the linguists, with the binary judgements provided for the associated training sets. In the experiments we use the first 10 of them (R101-R110). The TREC-11 training sets with binary judgments provided by the linguists are treated as the "perfect training sets" (named TREC sets), and would be used to compare with the training sets acquired from the Web using our proposed method (named ACTS sets). Two experimental models of *probabilistic* (named Prob) and *Ontology Mining model* (named OMM, see [7] for details) are developed with the modifications described above. They are used in the experiments to retrieve information from the RCV1 corpus. If the models trained by the ACTS sets can achieve the performance as approximately similar as (or close to) trained by the TREC

[2] Text REtrieval Conference, http://trec.nist.gov/.

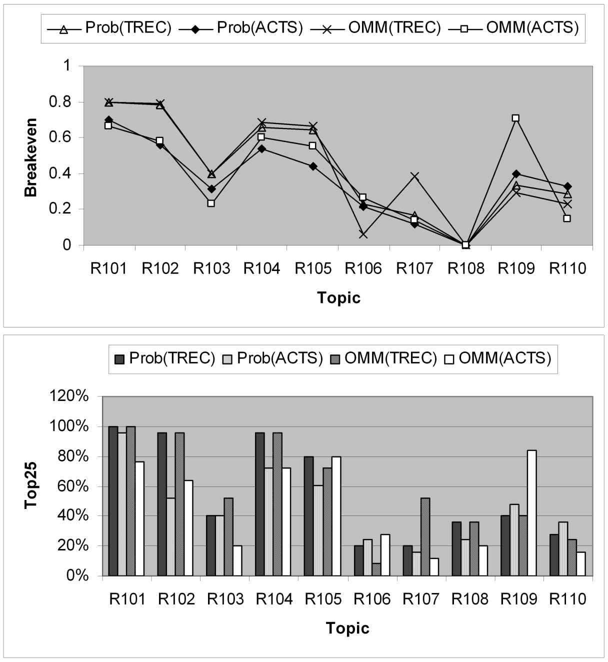

Figure 1. The experiment results

sets, the proposed method is proven to be successful. In order to measure the retrieval performance, $top\ 25\ precision$ and $breakeven\ point$ are chosen, which are two methods used in Web mining for testing effectiveness [7]. The greater both the top 25 precision and the breakeven point, the more effective the method is.

There are a total of 2016 Web documents gathered from the Web using the proposed method for the 10 experimental topics, 931 of them are identified positive, and 1085 negative. The comparison of the performances achieved by the two experimental models using different training sets is illustrated in Figure 1. One may see that the results based on the ACTS training sets are very close to the results based on the perfect TREC training sets. In some topics(R105, R106, R109, and R110), the results achieved by using the ACTS sets are even better than using the "perfect" TREC sets. According to the evaluation we believe that the proposed method is adequate to yield an promising achievement.

4 Related Works and Conclusions

Data classification and clustering have been used in Web log mining to discover new and interesting user behaviour patterns [8]. Association mining technique has been used in many Web usage mining systems to find correlations among Web pages and interesting access patterns [1] [11]. The technique has also been used for Web pre-fetching [12] and Web personalization [10].

In this paper we have presented an alternative method to automatically acquire training sets from Web data without user's heavy involvement. The method is developed based on the observations of the existing Web. The main contributions of the works are at first, a novel architecture for the effectiveness of using Web data, which saves training sets acquiring from expensive manual generation; secondly, a novel methodology of employing the observations of the existing Web to leverage Web information gathering, Web mining, and Web intelligence.

References

[1] T. Abraham and O. de Vel. Investigative profiling with computer forensic log data and association rules. In *Proc. of IEEE Int'l Conf. on Data Mining, Japan*, pages 11–18, 2002.

[2] B. G. Buchanan and H. Shortliffe. *Rule-based expert systems: The MYCIN experiments of the Stanford Heuristic Programming project*. Addison-Wesley, 1984.

[3] B. J. Jansen and A. Spink. How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Information Processing & Management*, 42(1):248–263, 2006.

[4] K. S. Jones. Information retrieval and artificial intelligence. *Artificial Intelligence*, 114(1-2):257–281, 1999.

[5] Y. Li, C. Zhang, and J. R. Swan. An information filtering model on the Web and its application in jobagent. *Knowledge-based Systems*, 13(5):285–296, 2000.

[6] Y. Li and N. Zhong. Ontology-based Web Mining Model: representations of user profiles. In *Proc. of IEEE/WIC Int'l Conf. on Web Intelligence*, pages 96–103, 2003.

[7] Y. Li and N. Zhong. Mining Ontology for Automatically Acquiring Web User Information Needs. *Knowledge-based Systems*, 18(4):554–568, 2006.

[8] B. Liu, Y. Ma, and P. S. Discovering business intelligence information by comparing company Web sites. In *N. Zhong and J. Liu and Y. Y. Yao (eds.) Web Intelligence*, pages 105–127. Springer-Verlag, 2003.

[9] Z. Y. Lu, Y. Y. Yao, and N. Zhong. Web log mining. In *N. Zhong, J. Liu and Y. Y.Yao (eds.) Web Intelligence*, pages 174–194. Springer-Verlag, 2003.

[10] B. Mobasher, H. Dai, T. Luo, Y. Sun, and J. Zhu. Integrating Web usage and content mining for more effective personalization. In *Proc. of Int'l Conf. on Ecommerce and Web Technologies, Greenwick, UK*, pages 165–176. Springer-Verlag, 2000.

[11] P. N. Tan and V. Kumar. Discovery of indirect associations from Web usage data. In *N. Zhong, J. Liu and Y. Y.Yao (eds.) Web Intelligence*, pages 128–152. Springer-Verlag, 2003.

[12] Q. Yang, H. Zhang, I. Tian, and Y. Li. Mining Web logs for prediction models in WWW caching and prefetching. In *Proc. of 7th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, pages 473–478, 2001.

[13] Y. Y. Yao, N. Zhong, J. Liu, and S. Ohsuga. Web Intelligence (WI) Research Challenges and Trends in the New Information Age. *Lecture Notes in Computer Science*, 2198:1, Jan 2001.