Topic-Oriented Information Detection and Scoring

Saike He, Xiaolong Zheng, Changli Zhang, and Lei Wang

Institute of Automation, Chinese Academy of Sciences, Beijing, China

Abstract. This paper introduces a new approach for topic-oriented information detection and scoring (TOIDS) based on a hybrid design: integrating characteristic word combination and self learning. Using the characteristic word combination approach, both related and unrelated words are involved to judge a webpage's relevance. To address the domain adaptation problem, our self learning technique utilizes historical information from characteristic word lexicon to facilitate detection. Empirical results indicate that the proposed approach outperforms benchmark systems, achieving higher precision. We also demonstrate that our approach can be easily adapted in different domains.

Keywords: characteristic word combination, Z-Score, self learning.

1 Introduction

TOIDS is a critical task in Intelligence and Security Informatics (ISI)[1,2]. Briefly, there are two problems demanding prompt solution. First, due to free wording and phrasing in web pages, most existing systems pale to differentiate topic related information from the others [3][4]. Secondly, traditional domain-oriented design blocks system's transferring ability, making domain adaption nontrivial [5][6]. In this paper, we propose a hybrid approach that utilizes characteristic word combination to improve domain-specific performance, while addressing the domain adaption problem by self learning.

The rest of the paper is structured as follows. Section 2 reviews related work in topic oriented information detection. In section 3, we present the architectural design and detailed technical information of our TOIDS system. Section 4 reports the results of our evaluation study. Section 5 concludes this paper with a summary.

2 Related Work

TOIDS has been attracting attention increasingly since the frequent occurrence of social security incidents [7]. Topics concerned in this paper include: pornographic violence, criminal offence, terrorism, public health and so on. Previous studies on topic-oriented information detection can be roughly categorized into two groups: heuristic dictionary-based methods and machine learning methods. In the following, we will review concrete research in each category.

2.1 Dictionary-Based Methods

Dictionary-based methods mainly employ a predefined dictionary and some hand generated rules for detecting topic-related webpage. Matching techniques, such as:

M. Chau et al. (Eds.): PAISI 2011, LNCS 6749, pp. 36–42, 2011.

Forward Maximum Matching (FMM), Reverse Directional Maximum Matching (RMM) and Bi-directional Maximum Matching (BMM) are employed to make relevance judgment based on word matching. Primal disadvantages of such systems lie in three aspects. First, the performance greatly depends on the coverage of the lexicon, which unfortunately may never be complete because new terms appear constantly. Secondly, without context taken into consideration, misunderstanding may be incurred in judgment procedure, especially when solid matching used over context sensitive words. Finally, judgment based on single characteristic word is highly unreliable, especially when several common words are used together to express topic related information[8].

2.2 Statistical and Machine Learning Methods

Some researchers cast topic-oriented information detection as a classification task. Li et al. [8] uses kernel based method to filter sensitive information. Greevy and Smeaton [3] classify racist texts with Support Vector Machine. In paper [7], Zhou et al. employ MDS to analyze hyperlink structures, thus uncovering hidden extremist group Web sites. Tsai and Chan [9] use probabilistic latent semantic analysis to detect keywords from cyber security weblogs. However, main drawback of such methods is their paleness in domain adaption, a key problem inherent in most statistical methods.

In this paper, we focus on improving detection precision by using characteristic word combination, akin to previous work in [8]. Apart from related words, we also involve unrelated words to make the final judgment. To facilitate domain adaptation, we integrate self learning technique by deducting based on historical prediction results. The following session will present all the technical details in our TOIDS system.

3 A Hybrid Approach for TOIDS

Figure 1 illustrated our proposed hybrid approach. We first extract characteristic words from training data, and then utilize topic related and unrelated characteristic words to generate word combination features for statistical model. Finally, we construct characteristic word lexicon with self learning based on prediction results. We believe this design could improve performance from two ways: word combination can help filter topic related web pages with higher accuracy. At the same, self learning will improve recall rate by enlarging its lexicon iteratively. This can also be considered as an effective way for domain adaption by utilizing information learned from new domain.

3.1 Z-Score Algorithm

Characteristic words originally refer to those words representative for topic related information. In order to extract such words, we employ Z-Score algorithm, similar to [10]. To meet our objective for TOIDS, we redefined the contingency tale, as given in Table 1.

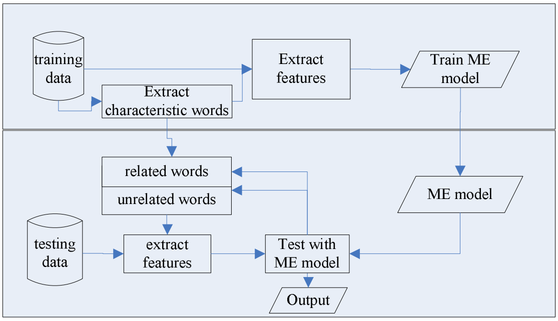

Fig. 1. Flow chart of TOIDS system

Table 1. Contingency table for characteristic word extraction

	Topic related	**Rest**	
ω	a	b	a + b
notω	c	d	c + d
	a + c	b + d	n = a + b + c + d

In table 1, the letter *a* represents the number of occurrences (tokens) of the word*ω* in topic related documents. The letter *b* denotes the number of tokens of the same word*ω*in the rest of the whole corpora while *a* + *b* is the total number of occurrences in the entire corpora. Similarly, *a* + *c* indicates the total number of tokens in topic related documents. The entire corpora corresponds to the union of ‘Topic related’ and ‘Rest’ that contains n tokens (n = a + b + c + d).

Here, we define a variable $Zscore(\omega)$ to measure the representative ability of word ω according to Muller's method [10].

$$Zscore(\omega) = \frac{\mathrm{a} - \mathrm{n}!\,\Pr(\omega)}{\sqrt{\mathrm{n}!\,\Pr(\omega)\cdot\left(1-\Pr(\omega)\right)}} \tag{1}$$

where:

$$\Pr(\omega) = (\mathrm{a}+\mathrm{b})/\mathrm{n} \tag{2}$$

$$\mathrm{n}! = \mathrm{a}+\mathrm{c} \tag{3}$$

As a rule, we consider words whose Z-Score values are above 0.8 as related while those below -0.8 as unrelated. The thresholds used here are chosen according to our best configuration in experiments. In word combination, we incorporate related and unrelated words into characteristic word set, for we believe both kinds of words are useful. With stop words pruned, related and unrelated words are added into the lexicon with their respective Z-Score value.

3.2 Statistical Model

As each webpage document is composed of sentences, thus, we intend to use the relevance of each sentence to derive a document's relevance, of which the former one comes down to a binary classification task. Here, we adopt Maximum Entropy Model (MEM) [11] to classify the relevance of each sentence. Features used here are mainly extracted based on characteristic words within a sentence, and grouped into four categories: n-gram word feature, n-gram POS feature, word number feature and finally, major POS feature. Table 2 lists these features with a detailed description.

Table 2. Features used in Maximum Entropy model

Feature	Type	Description
n-gram word	Related Unrelated	n-gram for related words n-gram for unrelated words
n-gram POS	Related Unrelated	n-gram for related POS tags n-gram for unrelated POS tags
word number	Related Unrelated	related word number in current sentence unrelated word number in current sentence
major POS	Related Unrelated	POS tag correspond to highest Z-Score value POS tag correspond to lowest Z-Score value

3.3 Topic Oriented Information Detection and Scoring Algorithm

To derive the relevance of document *d*, we introduce a variable *Rel_score(d)*:

$$Rel\_score(d) = \#\text{Rel\_Sentence}/\#\text{Sentence} \tag{4}$$

in which #Sentence indicates the total number of sentence in the current web page while #Rel_Sentence indicates the number of related ones predicated by machine. If *Rel_score(d)* exceeds 0.5, then it is considered as related.

For scoring with more concrete measurement, we also calculate the relevance degree for each web page. This is quantified by calculating the relevance probability in average for all sentences in a webpage, for we believe that prediction probability for each class could depict the relative importance of sentence in fine–grained level. In our TOIDS system, the original relevance degree is scaled into 5 levels[1].

3.4 Self Learning

For the problem of low coverage rate as well as domain adaption, we design a self learning technique by utilizing historical information. Specifically, we augment characteristic word lexicon based on prediction results: words that occur in related sentence yet not found in unrelated word lexicon are added to related word lexicon; words that occur in unrelated sentence yet not found in related word lexicon are added

[1] $Rel\_degree(d)$ values fall into span within [0, 0.5) are cast to 0, those within [0.5, 0.6) are cast to 1, those within [0.6, 0.7) are cast to 2, those within [0.7, 0.8) are cast to 3, and all above 0.8 are cast to 1.

to unrelated word lexicon. Thus, information from historical prediction can be reused in later classification phase.

4 Experiments and Results

We evaluate our approach from three aspects: the effectiveness of characteristic word combination, the influence of self learning and the ability of domain adaptation. Scoring results from actual system are also provided.

Corpora used in our experiments are comprised of about 5000 webpage documents from 10 websites, where 5 are non-governmental websites, the other are portals websites. One point worth mentioning here is that documents are crawled from websites directly rather than retrieval with specified key words, thus, the vocabulary set is open. This aspect guarantees that no priori knowledge of the test corpora is provided, which is highly similar to the circumstance in industrial applications. All collected web pages are manually labeled by professional knowledge engineers. Incompletely statistics show that there are about 20 percent of topic related documents in the whole corpora. The performance is measured by F-score, $F = (\beta + 1)RP/(\beta R + P)$, where R and P are recall and precision rate respectively. For high recall rate preference, we prioritize R by setting β to 2.

Table 3. System comparison under different configurations

Round	Precision	Recall	F-Score
ME + SingleRelWordMatch	62.38	84.36	75.49
ME + RelatedWordCom	79.84	82.75	81.76
ME + CharacteristicWordCom	82.30	82.81	82.64
ME + CharacteristicWordCom + Self-Learning	83.49	83.02	83.18

To test the effectiveness of characteristic word combination and self learning, we evaluate the performance of MEM under different configurations. In our experiment setting, 4-fold cross validations are conducted. Final results are reported in Table 3.

In table 3, 'ME + SingleRelWordMatch' serves as a base line system, which implements same combination strategy in [8]. In this experiment group, features for ME model are extracted only based on uni-gram related words. In comparison, round 'ME + RelatedWordCom' indicates that word combination could bring performance improvement. When unrelated words added to group 'ME + CharacteristicWordCom', performance was further boosted, from 81.76 to 82.64. This is primarily attributed to the improved detection precision, where unrelated words may play a key role. The effectiveness of self learning is justified from the last row in Table 3. For quantitative analysis, we have calculated ROOV value for our data set, which is defined as the ratio of word number in the training data compared to that in the testing data. The average value is 57.48, indicating a relatively low coverage rate. After classification procedure finished, statistics show that entries in our characteristic word lexicon increase averagely by 30 percent. We guess this portion of information remedies low coverage rate and ultimately attributed to overall performance improvement.

In domain adaptation experiment group, we subtract one sub-collection related to transportation (500 documents) to test another one concerning criminal incidents (400 documents). In this experiment, whenever one hundred new documents were classified, F-score is recalculated over all testing documents processed till the current time. Experiment results are given in Figure 2.

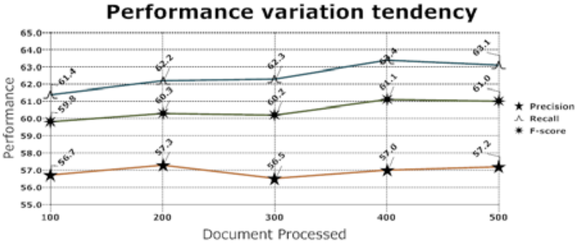

Fig. 2. Performance variation tendency

Table 4. Scoring results from TOIDS system

Title	Rel_score
一颗子弹 马英九连战矛盾面临 … One Bullet The contradiction between Ma Ying-jeou and Lien Chan faces …	5
男子为讨 68.6 万贷款持刀 … For debt collection of 686,000, men armed with knives …	4
河南交通厅长董永安落马… Transport Minister in Henan province Dong Yongan collapses …	5
被囚俄罗斯寡头能把 2 亿… Jailed Russian oligarch uses 200,000,000 …	3

Inspection into performance variation tendency in Figure 2 reveals that: historical prediction results have guided current classification phase. This consists with our original objective in design.

Apart from topic related information detection, results of our scoring strategy from a real TOIDS system are provided in Table 4, which only presents part of the filtered web pages with relevance degree level higher than 3.

5 Conclusions

In this paper, we propose a new method for TOIDS based on characteristic word combination and self learning. By integrating unrelated words into word combination, precision rate is improved while keeping recall rate at a high level. We solve domain adaptation problem by accumulate characteristic words from historical information.

Our future work includes the following:

- Use quantified Z-score of characteristic words to judge a sentence's relevance.
- Distinguish the importance of sentences occurring at different positions.

- Implement mutual enhancement mechanism between sentence and document for their relevance are prohibitively interdependent.
- Enable mistakenly extracted characteristic words eliminated automatically.

Acknowledgments

This work is supported in part by NNSFC #91024030, #60921061, #90924302, #71025001, #60875049 and #70890084, MOST #2009ZX10004-315 and #2008ZX10005-013, and CAS #2F070C01.

References

1. Wang, F.-y.: Social Computing: Fundamentals and Applications. In: Proceedings of IEEE International Conference on Intelligence and Security Informatics, ISI 2008, pp. 17–20 (2008)
2. Zeng, D., Wang, F.-y., Carley, K.M.: Guest Editor's Introduction: Social Computing. IEEE Intelligent Systems 22(5), 20–22 (2007)
3. Greevy, E., Smeaton, A.F.: Classifying racist texts using a support vector machine. In: Proceedings of the 27th Annual International ACM SIGIR Conference, New York, NY, USA, pp. 468–469 (2004)
4. Basilico, J., Hofmann, T.: Unifying collaborative and content-based filtering. In: Proceedings of International Conference of Machine Learning, New York, NY, USA (2004)
5. He, J., Liu, H.-y., Gong, Y.-z.: Techniques of improving filtering profiles in content filtering. Journal on Communications 25(3), 112–118 (2004)
6. Ma, L., Chen, Q.-x., Cai, L.-h.: An improved model for adaptive text information filtering. Journal of Computer Research and Development 42(1), 79–84 (2005)
7. Zhou, Y.-l., Reid, E., Qin, J.-l., Chen, H., Lai, G.: US domestic extremist groups on the Web: link and content analysis. IEEE Intelligent Systems 20(5), 41–51 (2005)
8. Li, W.-b., Sun, L., Nuo, M.-h., Wu, J.: Sensitive information filtering based on kernel method. Journal on Communications 29(4) (2008)
9. Tsai, F.S., Chan, K.L.: Detecting Cyber Security Threats in Weblogs Using Probabilistic Models, vol. 4430, pp. 46–57. Springer, Heidelberg (2007)
10. Zubaryeva, O., Savoy, J.: Opinion and Polarity Detection within Far-East Languages in NTCIR-7. In: Proceedings of NTCIR-7 Workshop Meeting, Tokyo, Japan, pp. 318–323 (2008)
11. Berger, A., Pietra, S.D., Pietra, V.D.: A Maximum Entropy Approach to Natural Language Processing. Computational Language 22(1) (2001)