DOI: 10.1051/itmconf/20171205018

Mining based on Extraction and Importance Evaluation Using Multi-Measures Methods for Electronic Documents

Wen XIONG, Zi-Hui DING

China Great Wall Computer Shenzhen Company Limited, China Electronics Corporation,Beijing, China
xiongwen@greatwall.com.cn

Abstract—Mining the implicit knowledge in the electronic documents is a critical task in text analysis and data mining. To attain a knowledge-based view of the electronic documents, the clustering method based upon the topic cannot only be used, but also that based upon the extraction can be done. Therefore, a novel method for the clustering of the electronic documents, summarizing of the full text based on the extracted segments, and an evaluation using multi-measures for the importance to the document were presented. In the method, eighteen kinds of named entities and two kinds of syntactical phrases were extracted, and exploited for the text clustering. Then, a novel similarity equation was proposed for the calculation about the extractions. Meantime, three measures for the importance to the document were proposed, which provided a different view for the document's content, and recommended a prior checking for the users. Therefore, the method can improve the efficiency of the knowledge discovery, and enhance the management of the document on the large scale of document collection.

1 Introduction

Along with the increases of the tremendous volume of the computer server, and the popularizing of the fast application of the cloud computing, many organizations have been accumulating more and more of the electronic documents. Mining the commercial value and the implied knowledge in the set of the electronic documents becomes an incremental market demand. The common methods in the text analysis and text mining include text classification, text clustering, named entity recognition (NER), information extraction (IE), text summary, document recommendation, etc.

From the angle of information retrieve (IR), a way of obtaining the information from a set of the documents is to index the full text of each document in the set based on words first, then, to use a search engine and some keywords to sort the relevant documents by calculating the similarity, and then, to return a sorted list to the user. The user will further examine the content of the documents, and locate the similar paragraphs. The result of the retrieve depends on the distribution of the keywords on the set of the documents.

Alternative methods from the text mining (TM), e.g. document clustering, calculate the document similarity according to the word morphologies and their distribution, which cluster the documents into some relatively independent classes according to the similarity rule, which can be used for analyzing the implied knowledge in the electronic documents. Both procedures above can be regarded as a pre-process, which are based on words, and extract the words' distribution, such as term frequency, term frequency and inverse document frequency (TF-IDF), etc.

However, there are many extractable segments instead of words, such as named entities (NE) and phrases, which represent the characteristic of the electronic documents, and their distributions in the documents represent the originality, and the quality. The finance relevant enterprises have a superior requirement to the active information, such as the stock, the events of corporation combination, the company's announcement, etc. E.g. the competitive companies care the commercial activity from the competitors, such as the fund invested to certain domains, the amount of money from the trade, and business show at certain localities in a special time. In addition, the jargons and phrases belong to the technology appeared at the interview of the competitors express the interests of the research and development (R&D) on the current or in the future.

Therefore, those named entities will be analyzed and extracted, such as company, location, people, time, money, number, and quantity. In addition, grammar chunks, such as the long verb phrases (VP) as the semantic of the activity, and the long noun phrases (NP) as the technical content of the R&D, will also be done.

Clustering and categorization for the documents often need to apply the feature reducing due to the high dimension of the word features in the document set. The common distribution of these words decides the inherent division of the document set. The result of the clustering only discloses the topic-level knowledge, which means the user cannot find the extraction-level knowledge intuitively from it, and the further process will be needed for text extraction. Therefore, the text extraction can be regarded as another pre-process different from that based on words mentioned above. Then, the text clustering and text classification based upon the extraction will provide other different viewpoints to the user, according to the kinds of the extractions and their similarity as the document features for the analysis, which exhibits the knowledge views of the document set, and therefore, have a special application value.

For the long documents, the summarizing for the full text provides a useful and a meaningful method. Generally speaking, it calculates the importance in the sentence based upon the importance in the words, then, it extracts the abstract by the importance in the sentence. Some named entities are ignored in the procedure, and time, number, and quantity are insensitivity. Therefore, it cannot satisfy the user's requirement from the angle of the IE, although this kind of summary can effectively shorten the content browsed by the user, and provide an abbreviate view of full text. On the other hand, the summarizing for the full text composed by the sentences based on the extraction segments will provide a new user view, which directly take the content extraction as the user's requirement, and form an important complement and an intuitionist expression for the abbreviated view of the full text, having a special application value.

Based on the above starting points, this paper presents a novel method for the clustering of the electronic documents, summarizing for full text based on extracted segments, and an evaluation using multi-measures for the document importance, which can be used for the analysis, mining, summary, evaluation, and recommendation for the electronic documents of the organizations.

The rest on the paper is organized as follows: first, in Section 2, the new method for mining and evaluation based on extraction is proposed; then, in Section 3, the related work in the mining and analysis of the documents is introduced in brief; in addition, in Section 4, the detail about the experiments is described; finally, in Section 5, conclusions are reached from the above discussion.

2 Methods

2.1Text Clustering Based on Extraction Information

To attain the cluster knowledge in the document set, we exploited the extracted information, such as named entities and grammar phrases. The former includes eighteen classes, such as Person, Facility, Geo-Political Entity (GPE), Product, Word of Art, Language, Time, Money, Ordinal, Nation or Region Politic (NORP), Organization, Location, Event, Law, Date, Percent, Quantity, and Cardinal. The latter includes two classes, i.e. long NP and long VP.

We can express the document as twenty variable-length vectors, which correspond to the eighteen classes of *NE*s and two classes of grammar phrases, and formulated the similarity of the documents as follows.

$$S(u,v)=\sum_{i=1}^{20}\max_{\substack{j=1,\dots,n1\\k=1,\dots,n2}}\{S^{'}(u_{ij},v_{ik})\}$$
$$s.t.\ S^{'}(u_{ij},v_{ik})=\begin{cases}|\,q_j=q_k\,|,\ i=Quantity\\ |\,c_j=c_k\,|,\ i=Money\\ |\,t_j\cong t_k\,|,\ i=Time,Date\\ u_{ij}==v_{ik},i=Ordinal,Cardianl,Percent\\ N-ed(u_{ij},v_{ik}),\ i=other.\end{cases}\tag{1}$$

Where: $ed(,)$ is the Levensthein edit-distance [1], and q for quantity, c for currency, t for time unit, and N for constant. Then, we adopted centroid-based clustering algorithms, e.g. K-Means to form the knowledge clusters.

2.2 Summarization for full text based on the adjoined position of the extracted information

Different from summarization for full text based on word importance, that based on the adjoined position of the extracted information needs to iterate each sub-clause of each sentence, and outputs the sentences and sub-clauses in the original order if they hit an extraction, which provides a view based on the extractions.

2.3 The importance evaluation for electronic documents using multi-measures integration

In the technical articles, a concept is more original when it was proposed early, or it was mentioned by the technical articles published after it. This typical phenomenon is appeared in the scientific papers and the patent texts [2]. Therefore, the paper proposes a measure for the originality of the documents, which reflects the original importance in the importance evaluation in the background set of documents as follows.

$$O_1=\sum\nolimits_{i=1}^{n}O(T_i)$$
$$s.t.\ \mathrm{O}(T)=max(\frac{sup(T)-2}{\text{age-in-days}(T)+1},0).\tag{2}$$

Where: $sup(T)$ is the support degree, and $\text{age-in-days}(T)$ is the number of the days when T appeared. The document is more important when the score of the document is bigger according to it.

The writing features are more complex when the number of the different words, named entities, long VP, and long NP is bigger. This typical phenomenon is also appeared in the scientific papers and patent texts [3]. Therefore, the paper presents a measure for the writing quality of the documents, which reflects the quality importance in the importance evaluation in the background set of documents as follows.

$$\begin{aligned} C &= IOWA_w(<v_1,a_1>,\ldots,<v_{10},a_{20}>) \\ &= \sum_{i=1}^{20} w_i a_{\sigma(i)} \\ s.t.\ & <v_1,\ldots,v_{20}> = <1,2,\ldots,20>. \end{aligned} \tag{3}$$

Where: $IOWA$ is the Induced Ordered Weighted Averaging operator, and w_i is the weight of a_i, corresponding to the twenty extractions; v_i is the induced variable. The document is more important when the score of the document is bigger according to it.

The concepts are often distributed widely when they are appeared in more articles in the background set of documents, which expresses they can be widely supported by other articles according to the extractions. Therefore, the paper proposes a measure for the concept distribution of the documents, which indicates the importance to the concept aspect in the importance evaluation in the background set of the documents as follows.

$$D_i = \sum_{\substack{i=1 \\ i\neq j}}^{n} S(u_i,u_j). \tag{4}$$

Where: u is the vector expression of the document mentioned at sub-Section II.A, and S is the similarity measure in the (1). The document is more important when the score of the document is bigger according to (4).

Meantime, the paper presents an integration evaluation using the *IOWA* operator on the above three measures as follows.

$$\begin{aligned} C &= IOWA_w(<u_1,b_1>,<u_2,b_2>,<u_3,b_3>) \\ &= \sum_{i=1}^{3} w_i b_{\sigma(i)} \\ s.t.\ & <u_1,u_2,u_3> = <3,2,1>. \end{aligned} \tag{5}$$

Where: the three measures are normalization according to the maximum and minimum scores, and b_i corresponds to the originality, the concept distribution, and the complexity scores of the document. w_i is weight of b_i, and u_i is the induced variable. Finally, the *C* value in the (5) is projected into the range of [0, 100] according to user's custom.

3 Related Work

Yu and Hsu [4] proposed a mining method based on content for the retrieve of the Computer-Aided Design (CAD) documents, which extracted the key phrases from the query and these documents according to the external knowledge base of the phrases, and modified the TF-IDF to express the weight of the key phrases, and to form the CAD-VSM (Vector Space Model). Then, the CAD document with maximum similarity of the query was returned, which indicated the effectiveness of the similarity calculation based on extracted information.

A dynamic semantic text clustering method was presented in [5], which used the extracted frequent items and named entities to express the document, and clustered the documents according to the expression. It reduced the dimensions of the document expression effectively due to the use of the extracted info.

To improve the clustering quality and the interpretation of the clustering results, a hierarchical clustering using *NE*s as privilege information was presented in [6], which can power the clustering solution. In our case, long *VP*s and long *NP*s are utilized besides the *NE*s.

An *NE* shared measure (NESM) was presented in [7], which was competitive to the standard similarity measure. Especially on the news documents, the *NESM* performed better than the standard measure.

4 Experiments

We carried out the experiments on the Chinese corpus[1], which needed the processes of the Chinese segmentation, the part of speech (POS), the named entity recognition, the component syntactic analysis, etc. We utilized the free ICTCLAS [8] tool due to its mature performance, and the free NiuParser [9] tool due to its sophisticated models based on the statistic, such as Conditional Random Fields, Average Perceptron, Maximum Entropy, and Recurrent Neural Network.

We executed a post-process on the results generated by the above tools, and extracted the useful knowledge. To illustrate the procedure of the post-process, we invented a commercial event with the correct expression as follows

> "ZaiXiangGangALiBaBaChuXiHuoDongDeALiBaBaGaoGuanWangShaoHuaYu 2017Nian2Yue28RiFanHuiBeiJing, BingDaiLaiLiao5QianWanYuanRenMinBiDe TouZiXiangMu, ZaiDaYue3000PingMiDeGongYeYuanQu, ZhuanLiFaMingJiShu JiangTongGuoTouZiZhuanHuaWeiShiJiDeGongYeChanPin."

Figure 1. Example of a commercial event.

After the Chinese segmentation and the POS tagging, the sentence was changed as follows.

[1] http://www.nlpir.org/download/tc-corpus-answer.rar

"Zai /p XiangGang /ns ALiBaBa /ns ChuXi /v HuoDong /vn De /u ALi /ns BaBaGao /a Guan /v Wang /nr ShaoHua /nr Yu /p 2017Nian /t 2Yue /t 28Ri /t FanHui /v BeiJing /ns , /w Bing /c DaiLai /v Liao /u 5Qian /m Wan /m Yuan /q RenMinBi /n De /u TouZi /vn XiangMu /n , /w Zai /p DaYue /d 3000 /m PingMi /q De /b GongYe /n YuanQu /n , /w ZhuanLi /n Fa Ming /vn JiShu /n Jiang /d TongGuo /p TouZi /v ZhuanHua /v Wei /v Shi Ji /a De /u GongYe /n ChanPin /n . /w"

Figure 2. Example after the component syntactic analysis.

We merged the adjoining words with the appropriate POS tags to form the extractions. The result was as follows.

Organization: ″XiangGangALiBaBa″ ; *Location*: ″BeiJing″ ; *GPE*: ″XiangGang, BeiJing″ ; *Person*: ″WangShaoHua″ ; *Quantity*: ″3000PingMi″ ; *Money*: ″5QianWanYuanRenMinBi″ ; *Date*:″2017Nian2Yue28Ri″.

Then, we utilized the NiuParser tool for the component syntactic analysis (CSA) to extract the long *NP*s, and the long *VP*s with a verb. The segmentations after the CSA were as follows.

(VP (VV ChuXi) (NP (NN HuoDong))),
(VP (VV FanHui) (NP (NR Beijing))),
(VP (VV DaiLai) (AS Liao) (NP (QP (CD 5QianWan) (CLP (M Yuan))) (NP (NN RenMinBi))),
(NP (NN TouZi) (NN XiangMu)),
(NP (NN GongYe) (NN YuanQu)),
(NP (NN ZhuanLi) (NN FaMing) (NN JiShu)),
(NP (NN GongYe) (NN ChanPin)).

Figure 3. Example after the component syntactic analysis.

Thereafter, we extracted the long *VP*s and the long *NP*s as follows.

Long *VP*s: ″ChuXiHuoDong, FanHuiBeiJing″ ; ″DaiLaiLiao5QianWanYuanRenMinBi″ ;

Long *NP*s: ″TouZiXiangMu, GongYeYuanQu″ ; ″ZhuanLiFaMingJiShu, GongYeChanPin″.

The number of the clusters in Sub-Section (II.A) can be given by the prior experience or by trial to obtain a satisfied effect (e.g. 5).

The weight vector for the twenty dimensions in (3), and the weight vector for the three dimensions in (5) can be obtained by the prior experience, or by training from the corpus tagged manually. Feasible values for the two vectors can be illustrated as follows.

$< w_1, w_2, ..., w_{20} > = < 0.05, 0.05, ..., 0.05 >$;

$< w_1, w_2, w_3 > = < 0.6, 0.3, 0.1 >$.

5 Conclusions

Mining the implied knowledge in the document set, we cannot only use the clustering method based on the topic, but also use that based upon the extraction. In addition, text summary based on the adjoining positions of the extractions was also used besides that based upon the sentence weight was done by use. The importance evaluation for electronic documents using multi-measures integration was adopted, which considered the document's originality, writing quality, and concept distribution. Those measures reflected the importance to the document by the views from the special angle, and an IOWA operator was used for integrating them into a final measure. In our internal testing, the earlier feedback from the users about the method was positive.

For future work, we will use the lexicon of the synonym for increasing the hitting score of the similarity between two documents, such as the abbreviation of the locations, and organizations. And more precise weights in the IOWA operators will be adopted by supervisor learning. Moreover, to benefit the model of the traditional topic-level, a hybrid similarity equation integrated with VSM of word weights and that of extractions with the extraction-level bias will be explored.

Acknowledgment

We used the ICTCLAS and the NiuParser for the Chinese segmentation, POS tagging, and syntactic parsing. Here, we express our thanks to the researchers for their mature open-source software, and our thanks to the reviewers for their review of the paper.

References

[1] V. I. Levenshtein, Binary Codes Capable of Correcting Deletions, Insertions and Reversals. Soviet Physics Doklady, 1966, 10:707-710.

[2] M. A. Hasan, W. S. Spangler, T. Griffin, and A. Alba, “COA: finding novel patents through text analysis,” in KDD’09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. New York, NY, USA: ACM, 2009, pp. 1175-1184.

[3] X. Jin, et al., "Patent Maintenance Recommendation with Patent Information Network Model." IEEE, International Conference on Data Mining IEEE, 2011: 280-289.

[4] Yu, Wen Der, and J. Y. Hsu. "Content-based text mining technique for retrieval of CAD documents." Automation in Construction 31.5 (2013):65-74.

[5] Yafooz, Wael M. S., et al., "Dynamic Semantic Textual Document Clustering Using Frequent Terms and Named Entity." IEEE, International Conference on System Engineering and Technology IEEE, 2013:336-340.

[6] Roberta A, Sinoara, et al., "Named entities as privileged information for hierarchical text clustering." International Database Engineering & Applications Symposium ACM, 2014:57-66.

[7] Montalvo, Soto, R. Martínez, and V. Fresno, "NESM: a named entity based proximity measure for multilingual news clustering." Procesamiento Del Lenguaje Natural 48(2012):81-88.

[8] H. P. Zhang, et al., "HHMM-based Chinese lexical analyzer ICTCLAS." Sighan Workshop on Chinese Language Processing Association for Computational Linguistics, 2003: pages. 758-759.

[9] J. Zhu, M. Zhu, Q. Wang, T. Xiao, Niuparser: A Chinese Syntactic and Semantic Parsing Toolkit, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing: System Demonstrations, 2015, pp. 145–150.