Research on Document Clustering from Internet Public Opinions

XimeiWang
Lanzhou University, Lanzhou730000, China;

93448597@qq.com

Keywords: topic discovery, clustering method

Abstract. Generally, in the traditional multilingual topic discovery, it is the multilingual text for a single goal of the conversion and then clustering. On this basis, we have constructed a custom dictionary for the people with the highest percentage of Chinese, Japanese and English in this paper. At the same time, we have improved the single-pass clustering algorithm in single language. And considering the characteristics of news effectiveness, we have proposed a multilingual text composite clustering algorithm based on fusion time impact factor, which makes the clustering analysis results more reasonable, and better reflects the characteristics of the effectiveness of network news.

1. Information acquisition pretreatment

Given the current, China's public interest in the main stream media language has the highest rate of Chinese, English and Japanese, while the English itself has been separated by a space between words, and in Chinese and Japanese, the character and the character is adjacent to, the connection between different words, the different connection between the same word can have different meanings. Therefore, how to divide the whole sentence correctly so that each character can be divided into the correct phrase, and then get the correct phrase segmentation, which is the focus of the multilingual information gathering and preprocessing.

1.1The proper noun thesaurus construction

Observation of Chinese and Japanese phrase structure, we will find that a large number of corpus will appear a large number of compound nouns,the commonly used word segmentation tools will separate these compound nouns into multiple nouns.The result is, although the identification of each phrase is accurate, ignoring that the original word is a compound noun. Obviously, it is not reasonable to have a long term to be divided into a compound noun, lost the meaning of the original word.

In order to solve the above problems, in this paper, the long term compound noun is used as a phrase to be added to the dictionary supported by word segmentation tools, thus, the word segmentation tool will not divide it into a number of phrases. Use the crawler technology of Baidu encyclopedia, the English Wikipedia, and Japan's version of Wikipedia to crawl the data, the process of crawling strategy is shown in Fig.1.

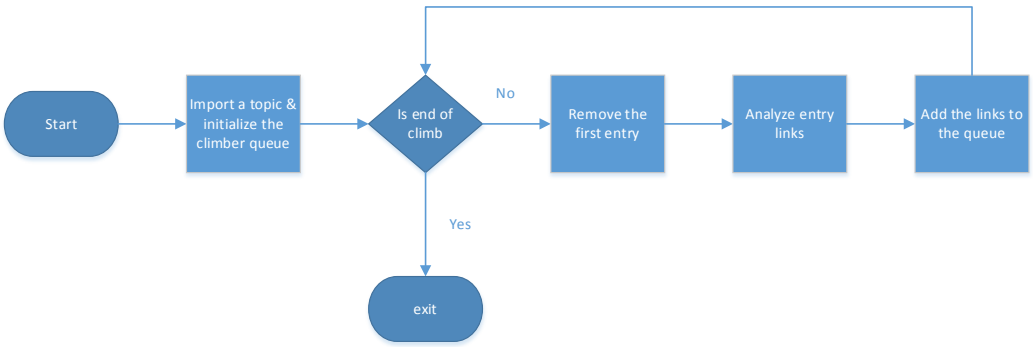

Fig. 1 Flow chart of the crawling strategy

1) Choose the keywords in the topic, and the words corresponding to Baidu encyclopedia and Wikipedia entry as the seed node of the climb queue of the web crawler;

2) Remove the first entry from the crawling queue, add the entry to the dictionary, analyze the entry links of all the Baidu encyclopedia and Wikipedia entries in the page, add the links which have not been climbed over to the climb queue;

3) Repeat the second step operation until the end of the climb is met or all entries are taken.

1.2 The disable thesaurus construction

After segmenting the news report and the Twitter short text data with the word segmentation tool, the TF-IDF algorithm is used to give a weight to each word, representing the importance of each word and the amount of information on the subject. While through the experiment of the actual data set, what can be found is that there are a lot of weight higher doesn't provide effective information for subject.

For example:

1) The common words on the Internet: reprint, original, thread, etc.

2) The meaningless words chosen from the clustering results: conflict, member and so on.

Analysis from the whole corpus, this kind of words appear not frequently, and not a lot of texts are containing these words. The IDFvalue of these words is also much higher than"I" these traditional stop words, leading to that these words which are not important have higher IDF, which is considered to be the corpus of the key words. Thus,in order to optimize the clustering results, on the basis of the traditional stop words, we have construct a disable thesaurus, which includes the keywords on this type of probability statistics.

2. Single language topic discovery

In the single language topic discovery, the single-pass Clustering algorithm is commonly used. It is simple and fast, and can meet the requirements of dynamic data processing. But, because of the special nature of news reports, the clustering accuracy is poor. Therefore, from the realization of clustering effect, we fully have explored the characteristics of the news reports, and improvedthe single-pass clustering algorithm specially based on the following aspects:

2.1 Topic structure adjustment

Subdivide the topic$T$ into various sub center topicas S_K, set the sub center topic similarity thresholdT_S, the topic of the new center thresholdT_N. In the process of clustering, the similarity between this report and the sub topic center of the topic is first compared, it is concluded that this report and the sub center topic S_K of the topic T have the largest similarity, and then compared with $T_S$andT_N.

2.2 In-depth analysis of the characteristics of network news reports

Introducing a weighting factor, increase the weight of the title word, person name, place name and organization name. Considering the title of the text is a high degree of content of the text, the text can be a good distinction between the categories, and therefore give a higher weight. In addition, the text of the name, place name, organization name and other characteristics of the title can also be a good area of the document category, which must be given a higher weight. The weight calculation formula of weighted factor is as follows:

$$w_i = \frac{tf_i(d)\log(\frac{N}{n_i}+0.01)}{\sqrt{\sum_{i=1}^{n}(tf_i(d))^2 * log^2(\frac{N}{n_i}+0.01)}} * F_k(3)$$

Where$F_k$is a weighting factor, which is used to adjust the weight of the name, organization name, place name, title, and feature word of the title, expressed as follows,

$$F_k = \begin{cases} \text{f1} \quad \text{The weight of the person name in the report} \\ \text{f2} \quad \text{The weight of the place names in the report} \\ \text{f3} \quad \text{The weight of the title of the report} \\ \text{f4} \quad \text{The weight of the organization's name in the report} \\ \text{f5} \quad \text{Other feature weight} \end{cases}$$

The content similarity between the text report$d$ and the topic$c$$\mathrm{sim(d,c)}$is calculated by the traditional angle cosine formula：

$$sim(d,c) = \frac{\sum_{j=1}^{n} W_{dj} * w_{cj}}{\sqrt{\sum_{j=1}^{n} w_{dj}^2 \sum_{j=1}^{n} w_{cj}^2}} (4)$$

2.3 If the time distance of these two reports is not considered, the single-pass algorithm often cluster them into the same topic, so the text time distance is introduced to further distinguish the document class, time distance calculation method is as follows:

$$dis(d,c) = \frac{2t_d - t_{ch} - t_{ce}}{2} (5)$$

Where$t_d$is the time to report the emergence of$d$, $t_{ch}$is the time for the first report of the topic$c$,$t_{ce}$is the time for the recent report of the topic$c$. The formula for calculating the similarity between the improved report $d$and the topic$c$ is as (6):

$$sim(d,c) = \alpha * \frac{\sum_{j=i}^{n} w_{dj} * w_{cj}}{\sqrt{\sum_{j=1}^{n} w_{dj}^2 \sum_{j=1}^{n} w_{cj}^2}} + \beta * \frac{2t_d - t_{ch} - t_{ce}}{2} (6)$$

The formula do not only consider the effect of the similarity based on the content, but also considerthe influence of the time factor.Among them，$\alpha + \beta = 1$，αdetermines the impact of the content similarity between the reports, andβdetermines the influence of the time distance. Because of the similarity of the contents of the report plays a supporting role, the factor of the time distance is in the auxiliary function, and thus$\alpha > \beta$. Through a large number of experiments, when $\alpha = 0.8$， Clustering effect is the best.

Based on the above analysis, the improved single-pass algorithm process is as follows:

1)Set the first report as the initial center of a topic.

2)Accept the next report$d$, calculate the weight of characteristics of each reportaccording to the formula (3).

3)Calculatethe similarity between$d$ andthe whole topic centers according to formula (6),through the comparison of the size,we find that d and a central $S_K$of thetopic c has the largest similarity P.

4)If$P > T_S$(sub center topic similarity threshold),then join d to the sub center $S_K$of the topic$c$.

5)IfT_N(topic new central threshold)$< P < T_S$, then get d as a new sub center of the topic c;

6)If the above 4) and 5)are not set up, then set d as a new topic of the initial center;

7)Clustering end, waiting for the next text vector.

3. The multilingual clustering based on fusion time factor

In the single language text clustering, according to the characteristic of the effectiveness of the news,the time characteristic is particularly important for the news text clustering. Learning from the experience of Journalism and Communication, the news will undergo initiation, development, climax, falling communication process from the published news to enter the quiet period.The earliest sources of traditional media coverage of the event statistical are 2.49 days in average, i.e., 60 hours event from exposure to enter the long tail of calm with an average of 29.64 days per play. Use Logistic function to describe the relationship between two post time interval Δt and the correlation between that publish short interval. The relation possibility of the news events is stronger and the time difference is in 14 days or so, the relation possibility of the news events begins to reduce, due to the long tail theory.Published in the interval long enough, it does not mean that two news reports would have no relevance, but it is more likely to be in low dependency conditions. Logistic function is as follows:

$$P(\Delta t) = \frac{a}{1 + Be^{-b\Delta t}} (7)$$

In the formula,Δt is time differenceof any two news reportspublished,a isweighting adjustment width of maximum distance between nodes.$P(\Delta t)$value in [0.5， 1.5], when the two reports published time close to, through the time weighting factor, the distance of the two news reports is closer.Otherwise,the distance is pushed away. This is not the case with a high cosine similarity in the original news text, but the similarity is lower, and the text distance is obviously more open, whichhelps the text clustering, and the boundary of thecluster is more clearly. According to the characteristics of news media reports, use Origin Pro9.0 to conduct Logistic function curve fitting,

curve parameters obtained:A_1=0.531 61，A_2=1.503 42，x_0=337.857 46，p=4.996 17. Function display, in 14 day time, the time factor adjustment function does not do the original results weighted, That is, the possibility that the larger correlation within two weeks of news reports, after that is significantly decreased with time.

4. Summary

In this paper, aimed at the present characteristics of the news media that the public concern, we select the most commonly used languages including Chinese, Japanese and English before translating them into a single cluster, which is a special vocabulary word and disables custom thesaurus. In single language clustering, improving single-pass clustering algorithm. At the same time, the time factor function is integrated in the complex multilingual text clustering algorithm for the timeliness features of news .Through the experiments, it is proved that the algorithm has better clustering effect, and better reflects the characteristics of the effectiveness of network news. However, the time complexity of the algorithm is increased after improvement. In addition, the title of the online news text is shorter and insufficient information, causing the deviation in the process of clustering. The next step of this work is to introduce of the latent semantic analysis (LSA) to further improve the efficiency of the algorithm in the title process.

References

[1]Lavrenko V, Allan J, DeGuzman E, et al. Relevance models for topic detection and tracking[C]: Morgan Kaufmann Publishers Inc., 2002: 115-121.

[2]QianLu.English and Chinese cross language topic detection and tracking technology research [D]: Minzu University of China, 2013

[3]PeiranZheng, DuoqianMiao, Zhifei Zhang, A method for the detection of Chinese micro blog news[J]. Computer science, 2012, (01):

[4]ChenghuiHuang, Jianyin, FangHou. A text similarity measurement method based on lexical entry semantic information and TF-IDF method [J]. Computer

[5]Lin G, Nagarajan C, Rajaraman R, et al. A general approach for incremental approximation and hierarchical clustering [J]. SIAM Journal on Computing, 2010, 39 (8): 3633-3669.

[6]Blei DM, Ng AY, Jordan MI. Latent dirichletallocation [J]. J Mach Learn Res, 2003, 3: 993-1