Automatic news summarization method based Maximal Marginal Relevance

Dengfeng Wei

Computer Science College, Yangtze University, Jingzhou 434023, China

weidengfeng@126.com

Abstract

Automatic summary, as the name suggests, is from a single article or multiple articles, the removal point to summarize the article to the effect of technology. It plays an important role in machine learning and data mining in order to obtain higher quality news summaries, news summaries proposed extraction method based on the theme of integration of mmr and gsvm. In news text recorded as a processing target to gsvm (support vector machine) and mmr (maximal marginal relevance) summary extraction algorithm is based around a topic for discussion for the current news, communication features to the theme as a basis for keyword scoring. Experimental results show that the proposed system when the system respectively svm summary, mmr and mmr summary system and svm combining summary comparison system, better extraction of summary effect of the former. This paper proposes a new sort of diversity learning model, by optimizing the evaluation of prospective edge for maximum correlation.

Keywords

Summarization; Sentence Similarity; Summary Extraction; Svm; MMR.

1. Introduction

With the rapid development of information technology and the Internet, the amount of information on the Internet to grow exponentially, and updated faster and faster, how to efficiently obtain useful information in the mass of information has become increasingly important. Abstracts As outlined text message content can be objectively summarize the main content of the information that allows people to efficiently obtain the required information through simple human-readable text. Abstract publisher of information resources, users and search engines are very important, many researchers summary method automatically get paid great attention.

Art automatic summarization news is that the most widely used, due to an overload of news and information, people are eager to have such a tool can help yourself with the shortest time to understand most of the most useful news. In addition, the search engine is one of the applications, query-based automatic summarization will help users find content of interest as soon as possible. The former is a single document summary technology, which is a multi-document summarization technology, the latter than in the former will be more complex [1]. Automatic summarization to solve the problem description is very simple, is to use some of the refined words to summarize the whole article to the effect the user can learn by reading the digest to the original meaning of the expression. Problem solving consists of two ideas, one is extractive, removable, find some key sentences from the original text, combined into a summary; the other is abstractive, Digest, which requires a computer can read the original content, and use their own means to express them. At this stage, the relative maturity of the program is removable, there are a lot of algorithms, there are some baseline testing, but is less desirable to give a summary of the study of the latter is not a lot of human language including

characters, words, phrases, sentences , paragraphs, document these level, research in ascending order of difficulty understanding sentences, paragraphs, yet difficult, not to mention the document, which is the biggest difficulty automatic summarization. Automatic summarization methods are mainly divided into two categories, extractive and abstractive. The former is the most mainstream, most applications, the easiest way, which is relatively more of a taste of true artificial intelligence. There is another classification method, single and multi-document summary document summary, the former is the basis of the latter, but the latter is not just the simple sum of the results of the former so simple[2].

2. GVSM and maximum edge correlation model

2.1 Section Headings

2.1.1 Sub-section Headings

We have a "common vocabulary" hypothesis: "Documents" and "inquiry", equivalent to a collection of words they contain, their relevance can be completely by the case of a total vocabulary to determine the vector space model, the most simple binary vector just a word appears portray item or not, a little more complicated, count vector, depicts the number of items in a word document appears, generally, we can consider a document set as the background, the right to a lexical item in a document is important MMR one kind of redefining how-order values. Jth Document is a unit of text index, for example, a web page, a news story, an article, a patent, a legal case, a word, a word, a word, A book, a chapter of a book, etc. Tf (wi, Dj) is the "Term Frequency:", word frequency is the number of wi appear in the document Dj in. People sometimes by dividing the document in the largest non-stop words of TF, Tf to normalize [Tf norm = Tf / max_TF].

$$\max\_TF(D_j) = \max_{w_i \in D_j}(TF(w_i, D_j)) \tag{1}$$

Df (w_i , C) is "document frequency , w_i At least one of the number of the document appears in them. Usually we take the results normalized, ie divided by the total number of documents in C. Idf (w_i , C) is "Inverse Document Frequency", [Df (w_i , C) / size (C)] - 1. In most cases people use Log_2 (I D_j), Rather than directly IDf. Words key phase in TfIDf sense for a document weight, in general: TfIDf (w_i , D_j , C) = F1 (Tf (wi, D_j) * F2 (IDf (w_i , C)). Typically, F1 = 0.5 + Log_2 (Tf), or Tf / Tfmax or 0.5 + 0.5Tf / Tfmax. Typically, F2 = Log_2 (Idf), "suppressing function" in Salton's SMART IR system: TfIDf (w_i , D_j , C) = [0.5 + 0.5Tf (w_i , D_j / Tfmax (D_j)] * Log_2 (Idf (w_i , C)). The document collection is expressed as a vector: Let C = [D1, D2, ..., Dm] each lexical item in accordance with its distribution in the collection of documents is also represented as a vector: Let vec (ti) = [Tf (ti, D1), Tf (ti, D2), ..., Tf (ti, Dm)]. The similarity between the definition of lexical items:

$$sim(t_i, t_j) = \cos(vec(t_i), vec(t_j)) \tag{2}$$

Thus, the term is often occur simultaneously would be higher similarity. Query-Document similarity calculation corresponding change, sim (q, d) no longer q and d vector dot product, but with the above "Terms - Terms" in a function similarity. For example, for every word entry q respectively obtain maximum likelihood it and d lexical items, the similarity of these add up to the maximum similarity of q and d:

$$sim(q,d) = \sum_i [maxj(sim(tqi,\ tdj)] \tag{3}$$

Usually the length q and d on the basis of normalized do:

$$simnorm(Q,\ D) = \frac{\sum[Max(sim(q_i, d_j)]}{|Q| \times |D|} \tag{4}$$

2.2 Maximum Edge Related.

$$MMR = \arg\max_{D_i \in RS} \left[\lambda Sim_1(D_i, Q) - (1-\lambda) \max_{D_i \in RS} Sim_2(D_i, D_j) \right] \quad (5)$$

Di: Documents in the collection C

Q: Query

R: Relevant documents in C,

S: Current result set

Assume that we are given a database of 5 documents di and a query q, and we calculated,Given a symmetrical similarity measure, the similarity values as below. Further assume that Is given by the user to be 0.5.

S =

	d_1	d_2	d_3	d_4	d_5	q
d_1	1	0.11	0.23	0.76	0.25	0.91
d_2		1	0.29	0.57	0.51	0.90
d_3			1	0.02	0.20	0.50
d_4				1	0.33	0.06
d_5					1	0.63
q						1

Fig. 1 Maximum Edge Related

Currently our result set S is empty. Therefore the second half of the equation, which is the max pairwise similarity within S, will be zero. For the first iteration, MMR equation reduces to:MMR = arg max (Sim (di, q)),D1 has the maximum similarity with q, therefore We pick it and add it to S. Now, S = {d1 }.

The use of MMR performed a computational method for document reordering

Step 1. Made before K documents with other commonly used method of IR

Note Dr = IR (C, Q, K)

Step 2. Election max sim Ranked = <di>

Step 3. Let Dr = Dr, which remove this element

Step 4. While Dr is not empty, do:

A. Find di with max MMR (Q, Dr, Ranked)

B. Let Ranked = Ranked • di, c. Let Dr = Dr

As can be seen from the above equation sim (Q, di) it represents di relevance, and sim (di, dj) di represents redundancy; the core of MMR, ie weigh these two properties, namely, redundancy = cost, relevance = benefit

It uses a maximum edge correlation model (Maximal Marginal Relevance) a variant. MMR is unsupervised learning model, it was proposed to improve the performance of information retrieval (Information Retrieval) system. For example, a search engine that will be most frequently used information retrieval system. You may often encounter, for we enter a keyword, the search engine will usually give duplicate content or too close to the case retrieval. To avoid this phenomenon, the search engine may be increased by MMR diversity of content, search results are given various considerations, in order to improve performance. Because it is in line with a summary of the basic requirements that weigh relevance and diversity. Understandably, more relevant results and a summary of the original, it is close to the center full meaning. Diversity is considering making more comprehensive summary of the contents. Very intuitive and simple is an advantage of the model. Compared to other learning methods unsupervised, as TextRank (TR), PageRank (PR) and so on, MMR is considering the diversity of information to avoid duplication results. TR, PR is a diagram (Graph) learning method

based on each sentence as points, two points between each band has a heavy weight (Weighted) undirected edges. The weight of the edge implicitly defines the probability of travel between different sentences. These methods make the summary of the problem as a random walk to identify high probability distribution at steady state (Stable Distribution) under the (important) set of sentences, but one can not avoid a similar drawback is elected sentences between each other a very high degree of phenomenon.

The MMR method can solve the problem of diversity of choice sentences. Specifically, in the MMR model, while the relevance and diversity measure. Thus, you can easily adjust the relevant rights and diversity to meet the heavy bias "require similar content" or bias "require different aspects of" requirements. For relevance and diversity of specific assessment, it is defined by semantic similarity between sentences to achieve. Sentence similarity is higher, the more relevant and lower diversity.

In order to obtain the initial k documents, you can use other relatively simple information retrieval (IR), such as the common law, sub-laws, retroactive legislation, etc., thus obtaining the K beginning of the document, that is the total set of documents; then from Select the Query closest to a document, marked the first document, and then removed from the K documents, the set as ordered, namely R; and documentation set for all documents using MMR formula to find such MMR the maximum document, added to the ordered set.

And so forth, to re-determine the order value of the document. MMR formula we need to adjust the parameters k and lamda. Removable method based on an assumption that the core idea of a document can be a document or a few words to summarize. Then the task becomes a summary of the most important documents to find the words, which is a sort of problem. Sorting is a very classic issue, but also a lot of solutions to problems. For example: Google page list according to the user's query generated, the result is a sort after; Another example is Amazon's recommendation system recommended to the user of the N may be interested in the product, are also made by the algorithm to sort the output. For different sort of problem, need to make different indicators, such as some applications are concerned with relevance, timeliness is of some concern, some concern is the novelty and so on, up the discussion at this level sorting, will be different model. Removable general summaries, will consider the relevance and novelty of the two indicators. Relevance refers to the summary sentences used in this document represents the best possible means, and novelty means that the redundant information contained in the candidate sentence less, as every word can be independently express an independent meaning.

3. MMR in text summarization

Summarization tasks can be considered as a binary classification problem, and you can use supervised learning methods.To solve, in a supervised approach, each training and testing is through a set of sentences indicative of the features.That the use of positive or negative label to represent whether a sentence is a summary sentence. This chapter uses the support vector machine .SVM sentence for each test data set, by predicting the sentence confidence value to determine whether the sentence contains.Among the summary. According to a certain compression ratio to select the sentences with a high confidence value, thereby forming the summary. In A news, the number of summary sentences are usually far less than the number of non-sentence summary, so as SVM classifier will.There is a data imbalance on SVM hyperplane some confidence value similar sentence is actually very difficult to judge,Off if it is a summary sentence, then the data in order to solve the problem of unbalanced, using unsupervised learning methods largest edge Margin-Related (MMR) method of extracting news summaries, which is a digest without a lot of manual annotation.

Extraction method, by calculating the sentence MMR values, and then follow the MMR sentence summary values sorting, and press .According to the level of worth to make further screening summary sentence, but according to this method to extract a summary sentence did not take into

Position information of the sentence, it is also necessary to select the summary sentences are ordered according to their position in the information document.

However, supervised learning method requires a lot of manual annotation, time-consuming, but a summary in accordance with the level of confidence value will be Screening, resulting in a final summary does not logically structured. Follow-Up treatment is also required in accordance with the sentence in the news text Sentence position corresponding sort, and then get the most accurate summary sentence.

3.1 News text feature selection and pretreatment.

Among a press release, the press start position and end position of the sentence for the summary sentence more likely in the following figure. Sentence Sub-Position weight scale factor, usually set according to their actual situation, generally the initial segment of the sentence given to a weight ratio of paragraph 2.5, the end of the paragraph the sentence given its proportional weight of paragraph 2. In addition, some of the sentence is a sentence contains a special word, these important sentence is generally "In conclusion", "In short", etc., but those sentences as unimportant "It is impossible", "There is nothing that" Wait. These sentences are a great contribution to the news summary.

First, extract various feature and the feature word processing. Typically, in a press release, the role of the largest general noun or verb, noun and verb and therefore special handling during tagging, in addition to some pauses words such as "well, uh, um" While the use of the words a high rate, but does not reflect News by topic, from time to time in the label given to its lower weight. Extract a series of related features, this paper focuses on text messages without using the news will or rhythmic features.

3.2 Lexical Features.

The following table lists the lexical features need to be extracted, the first column contains the number of sentence length and word of each sentence which, these contents are removed after a pause word acquired. According to the law the sentence summary researchers concluded that under normal circumstances, a longer sentence, the information it contains the more, then the likelihood of becoming a summary sentence greater. Therefore, the extracted feature includes a sentence before the current sentence and the next sentence length information. The study also use the 'Frequency' and 'two yuan grammar Frequency', these features indicate the frequency of automatically generated sentence, one-gram grammar and binary words, the frequency of these words in the sentence other than the frequency of the word appears to be high. The study showed that a sentence in the first occurrence of nouns and verbs tend to offer very important information, in order to extract features also include the term first appeared in a sentence or the number of verbs. The following table lists the feature information extracted vocabulary.

Table 1. Lexical feature information

Numble	Scheme 1
Feature	Feature description
Len I, II, III	The length of the previous sentence, the current sentence and the next sentence
Num I, II, III Unigram	The number of words in the previous sentence, the current sentence and the next sentence Word frequency

3.3 Characteristics topic.

News speech contains a lot of stories like fragments, they are still divided into several parts, and each part has its own theme. In the news summary extraction among topics fragment also contains useful information. In order to better achieve different topics related characteristics, this study contains a number of themes related features. The subject of the study related to the use of the feature is based on the theme of the so-called term frequency (TTF) and theme switching frequency (ITF), the calculation

of these two features is based on the basis of information recorded on the topic. TTF term frequency is just a topic in which ITF worth calculation method 6.

Equation:

$$ITF(w_j) = \log(NT / NT_j) \quad (6)$$

Nti which is the topic that contains the number of words in a news ùi, NT is the number of topics in the news segment. Where ti is the right word word ùi weight calculation TF-IDF values for use. Table 2 under the theme related feature extraction.

Table 2. Three Scheme comparing

Feature	Feature description
ITF I, II, III	ITF average, maximum and value sum
TTFITF I, II, III	TTF * ITF average, maximum and value sum
Feature	Feature description

4. The Simulation Results and Analysis

4.1 Experimental Data

The experimental data used in this article: 1 from Netease news data interface 2015-2016 500 News; 2 to extract 500 comprising tech from Sina, Baidu, Tencent, and other sites in the End of the World, education, culture, military affairs in length News in the range of 500-10000 word; 3. Western media Collections News 2014 data, 2015 Winter Olympics, World Cup, the 2015 US mid-term elections, the movement of the New York Times published in Sochi news analysis, when the introduction of the World Cup interactive news readers, Washington Post during the launch of the US mid-term elections 2014 election Visual Analysis series news.

4.2 Evaluation

It presented a summary of the features and a summary measure of the quality and characteristics of the weighting, and the experiment of Chapter III proved to be effective. Thus, in the experiment of this chapter, not only the use of evaluation based on artificial and acceptability based on a reference Summary Evaluation two evaluation methods, and a summary of feature weighting and also as a method to evaluate the experiment, feature weighting and calculation formulas summary is as follows:

$$Score(x) = w_1 * f_1(x) + w_2 * f_2(x) + w_3 * f_3(x) \quad (11)$$

F1 represents the cost summary explanation, see the formula above formula, f2 represents redundancy digest value, f3 represents the relative length digest, w1, w2, w3, respectively document f1, f2, f3 weight. Through these three evaluation methods for a variety of algorithms and other algorithms presented in this chapter experimental analysis and comparison.

5. Conclusion

News Summary extraction is an important field of artificial intelligence, text retrieval, has been the concern, also raised a lot of ways. This paper documents the first word, and then stored in the graph structure, and then extract keywords, and then use the database connection between words is generated sentence summary. This article is only for keywords and text of article abstracts were extracted and the effect of artificial summary results are basically consistent. After the text is available on-line database to improve the accuracy and timeliness of summary. The method of automated news summary process presents a new idea, want to delve into this area to make that contribution. In the next step, the further study and improve this method to do it.

References

[1] W. JI, Z. LI, W. CHAO and X. CHEN, "A New Method for Calculating Similarity between Sentences and Application on Automatic Abstracting," Intelligent Information Management, Vol. 1 No. 1, 2009, pp. 36-42. doi: 10.4236/iim.2009.11007.

[2] J. Motta, L. Capus and N. Tourigny, "Insertion of Ontological Knowledge to Improve Automatic Summarization Extraction Methods," Journal of Intelligent Learning Systems and Applications, Vol. 3 No. 3, 2011, pp. 131-138. doi: 10.4236/jilsa.2011.33015.

[3] K. S. Jones, "Automatic Summarising: The State of Art," Information Processing and Management, Vol. 43, No. 6, 2007, pp. 1449-1481. doi:10.1016/j.ipm.2007.03.009.

[4] J. Goldstein, "Evaluating and generating summaries using normalized probabilities," Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, 1999, pp. 121- 128. doi:10.1145/312624.312665

[5] K. S. Jones, "Automatic Summarising: The State of Art," Information Processing and Management, Vol. 43, No. 6, 2007, pp. 1449-1481. doi:10.1016/j.ipm.2007.03.009

[6] T. A. Lasko, J. G. Bhagwat, K. H. Zou and L. Ohno-Machado, "The Use of Receiver Operating Characteristic Curves in Biomedical Informatics," Journal of Biomedical Informatics, Vol. 38, No. 5, 2005, pp. 404-415. doi:10.1016/j.jbi.2005.02.008

[7] R. Alguliev, R. Aliguliyev and M. Hajirahimova, "Multi-Document Summarization Model Based on Integer Linear Programming," Intelligent Control and Automation, Vol. 1 No. 2, 2010, pp. 105-111. doi: 10.4236/ica.2010.12012.

[8] C. Sitaula and Y. Ojha, "Semantic Sentence Similarity Using Finite State Machine," Intelligent Information Management, Vol. 5 No. 6, 2013, pp. 171-174. doi: 10.4236/iim.2013.56018.

[9] I. Beltagy, et al., "Montague Meets Markov: Deep Semantics with Probabilistic Logical Form," 2nd Joint Conference on Lexical and Computational Semantics: Proceeding of the Main Conference and the Shared Task, Atlanta, 13-14 June 2013, pp. 11-21.

[10] Wang, R. , Wang, C. , Xu, Y. and Cui, X. (2016) The Research of Chinese Words Semantic Similarity Calculation with Multi-Information. International Journal of Intelligence Science, 6, 17-28. doi: 10.4236/ijis.2016.63003.

[11] M. Mehr, "Using AdaBoost Meta-Learning Algorithm for Medical News Multi-Document Summarization," Intelligent Information Management, Vol. 5 No. 6, 2013, pp. 182-190. doi: 10.4236/iim.2013.56020.