Commodity Attribute Extraction

NLP: Phrase Mining, Keyword Extraction

Menu

Commodity Attribute Extraction........ 1
NLP: Phrase Mining, Keyword Extraction........1
1 Definition........ 1
2 Problem Analysis........ 2
3 Unsupervised algorithm........2
3.1 Based on statistical features (TF, TF-IDF)........ 2
3.2 Based on word graph model (PageRank, textrank)........2
3.3 Based on topic model (LDA)........2
3.4 Automatic phrase mining framework--Autophrase........2
3.5 Keyword matching........3
3.6 Word2vec word clustering (DBSCAN / K-means)........ 3
4 Supervised algorithm........ 3
4.1NER........3
4.2 Based on information gain........3
4.3 Key words extraction of chi square test........3
4.4 Convert to multi-label problem (manual tagging, training model)........3
4.5 CRF........ 3
4.6 Part of speech + rules........4
5 Conclusion........4
6 Reference........5

1 Definition

Commodity attribute words generally belong to the key words of reviews, so the extraction method of attribute words is basically the same as that of key words. Key words are words that can express the content of the document center. They are often used in the computer system to index the content characteristics of the paper, information retrieval, system collection for readers to review. Keyword extraction is a branch of text mining, which is the basic work of text mining, such as text retrieval, document comparison, summary generation, document classification and clustering.

2 Problem Analysis

From the perspective of algorithm, there are two kinds of keyword extraction algorithms: unsupervised keyword extraction and supervised keyword extraction.

3 Unsupervised algorithm

3.1 Based on statistical features (TF, TF-IDF)

Idea: extract key words from documents by using statistical information of words in documents;

Methods: to reflect the importance of a word to a document, filter out common words, and retain important words;

Advantages and disadvantages: strong generalization ability, no need for manual annotation.

3.2 Based on word graph model (PageRank, textrank)

Idea: first, we should build the language network diagram of the document, then analyze the language network diagram, and find the words or phrases that have important functions on the diagram, which are the key words of the document;

Methods: the words in the text were regarded as the nodes in the graph, and the nodes with high weight were regarded as the key words through the connection of edges;

Advantages and disadvantages: most of the words extracted are high frequency words, which have limitations;

3.3 Based on topic model (LDA)

LDA, also known as three-tier Bayesian probability model, includes three-tier structure of words, topics and documents. It clusters words by topics by using the co-occurrence relationship of words in documents, and obtains two probability distributions of "document topic" and "topic word".

Idea: to extract key words by using the properties of topic distribution in topic model;

Methods: segmentation **-** > de stop words **-** > building word bag model **-** > LDA model training

Advantages and disadvantages: the document is too short to train LDA; the relationship between words is not considered;

3.4 Automatic phrase mining framework--Autophrase

Method: extract phrases from autophrase + manual review results

Advantages and disadvantages: Based on Wikipedia scene; high extraction accuracy and low labor cost

3.5 Keyword matching

Methods: the corresponding text was matched directly according to the product attribute lexicon;

3.6 Word2vec word clustering (DBSCAN / K-means)

Idea: for words represented by word vectors, the DBSCAN / k-means algorithm is used to cluster the words in the article. The clustering center is selected as a main keyword of the text, and the distance between other words and the clustering center is calculated, that is, the similarity. The top k words closest to the clustering center are selected as the keywords, and the similarity between words can be calculated by the vectors generated by word2vec.

Methods: word segmentation > word2vec > DBSCAN / K-means clustering

4 Supervised algorithm

4.1NER

Idea: train ner model according to existing product attribute table

Method: train ner with {comment, commodity attribute}

Advantages and disadvantages: need a lot of manual annotation data to train the model;

4.2 Based on information gain

4.3 Key words extraction of chi square test

Chi square is a method used to test the independence of two variables in mathematical statistics. It is a statistical method to determine whether there is correlation between two classified variables. The classical chi square test is to test the correlation between qualitative independent variables and qualitative dependent variables.

Idea: select features by chi square test according to annotation data

4.4 Convert to multi-label problem (manual tagging, training model)

4.5 CRF

Conditional random fields (CRF) is a conditional probability distribution model for a given set of input sequences and another set of output sequences. For example, in the part of speech tagging task, the input sequence is a string of words, and the output sequence is the corresponding part of speech.

Methods: three features are usually selected: the front and back position words of attributive words as features; the core words and sentences are extracted by syntactic analysis, and the syntactic structure is taken as features; the part of speech and the word spacing between attributive words and emotional words are taken as features. Training CRF model

4.6 Part of speech + rules

Idea: according to the established rules, we can divide the sentence into words, mark the part of speech, and then extract nouns and adjective gerunds and so on.

Methods: segmentation, part of speech tagging and rule extraction

Advantages and disadvantages: simple and fast; disadvantages: low accuracy, mostly used for tag extraction

5 Conclusion

Text keyword extraction is widely used in text-based search, recommendation and data mining. At the same time, in practical application, because of the complexity of application environment, for different types of text, such as long text and short text, the same text keyword extraction method has the same effect. Therefore, in the practical application, the algorithm used for different conditions will be different, no one kind of algorithm has a good effect in all environments.

Compared with the single algorithm mentioned above, some combination algorithms are widely used in engineering to make up for the shortcomings of the single algorithm, such as combining TF-IDF algorithm with textrank algorithm, or synthesizing TF-IDF and part of speech to get keywords. At the same time, there is a great dependence on text preprocessing and the accuracy of text segmentation in engineering. For the wrong words, deformed words and other information of the text, it needs to be solved in the pre-processing stage. The choice of word segmentation algorithm, the recognition of unknown words and ambiguous words will have a great impact on keyword extraction to a certain extent.

6 Reference

[1] Liu, Xiaoming. Domain-Specific Ontology Construction from Hierarchy Web Documents[R]. Beijing:IEEE, 270 2011. 160-163.

[2] Yanhui, Zhu. Research on Extraction for Feature Words from Chinese Product Comments[R]. Changsha:IEEE, 2010. 205-209.

[3] 任远远，王卫平．中文网络评论的产品特征提取及情感倾向判定[J]. 计算机系统应用,2014,(11):22-27.DOI:10.3969/j.issn.1003-3254.2014.11.005.

[4]http://manu44.magtech.com.cn/Jwk_infotech_wk3/article/2011/1003-3513/1003-3513-27-5-49.html#outline_anchor_11

[5]https://blog.csdn.net/asialee_bird/article/details/96454544#%E5%8D%81%E3%80%81%E6%80%BB%E7%BB%93

[6] https://zhuanlan.zhihu.com/p/40631031

[7] https://blog.csdn.net/Sakura55/article/details/85122966