TEXT MINING AND EXPLORATION USING SVM

[1]Arun.k, [2]Syed Mohammed Shafi , [3]J Sushma , [4]Veena Rani
[1,2,3,4] Assistant Professor
Department of Computer Science and Engineering,
Malla Reddy College of Engineering, Hyderabad

Abstract

On the basis of analyzing the basic concepts and the process of text excavation, the present study proposes some new methods in extraction of text features, deflation of characteristic collection, extraction of study and knowledge pattern, and appraisal of model quality. Meanwhile, it makes a comparison of two types of text categorization, text classifications and text cluster, and it briefly explores the basic issues to be solved in the future development of the text excavation technology.

Key words- Text excavation; Text Features; characteristic collection deflation; Text classification; Text cluster

Ⅰ. INTRODUCTION

Along with the Internet application's popularization, Web already developed into has 300,000,000 page's distributional information spaces, moreover this digit still by the speed which every half year doubles grew. In the middle of these mass data, the majority of information are the non structuriation perhaps half structurization, moreover is containing the huge potential value knowledge. The people urgent need can from Web fast, discover these valuable knowledge effectively. On Web the information multiplicity has decided the Web knowledge discovery multiplicity. According to the processing object's difference, may the Web knowledge discover that divides into two broad headings: Content discovery and structure discovery. The content discovered that is mainly the excavation which keeps off to article this article. The text excavation (Text Mining), may the massive documents set content carry on the abstract, classified, the cluster, the connection analysis as well as to Web on carries on the tendency forecast to the documents and soon.

Ⅱ. BASIC CONCEPTS

The text is by the massive characters, the word, the sentence is composed, to text excavation, in paramount consideration text character word. In English, Chinese and so on the natural language, have the massive words the concurrently kind of phenomenon, this for the text part-of- speech tagging, semantic labeling has brought the very major difficulty. Therefore, how to remove the part of speech, the semantic different meanings, is the text automatic labeling research key question.

A. A part-of-speech tagging

Ⅰ) Concurrently kind of word: Has two or two above lexical category glossary calls the concurrently kind of word.

the concurrently kind of word displays the different semantics in the different context linguistic environment, is by the concurrently kind of word lexical category decided that this is in the text excavation part-of-speech tagging question.

Concurrently kind of word classification Same-type opposite sex different righteousness concurrently kind of word

For example: Chairman Mao leads us to fight for state power. ("leadership" is a verb, leads, meaning of the instruction)

Chairman Mao is our good leadership. ("leadership" of is noun, meaning of person in charge, the leader)

Same-type opposite sex synonymy concurrently kind of word

For example: He has worked for 3 hours.

(“hour” is classifier, Unit of time)

We measure the operating time by the hour. (“hour” is noun, Unit of time)

Heterogeneous homogeneous synonymy concurrently kind of word

For example: The computer has bought 50 computers. (“computer” is noun and “computer” synonymy)

The computer has bought 50 computers. (“computer” is noun and “computer” synonymy)

The non-word usage (stops word usage): In text relatively auxiliary functional word.

Ⅱ)Non-word usage classification

Function word: In English “a, the, for, with,…”; In Chinese “,…”And so on.

Full word: In database conference's paper “database” a word, although the frequency of use is very high, but regards as the non word usage.

Ⅲ) Stem question: compute, computes, computed identifies a word (distortion).

Ⅳ) part-of-speech tagging: The so-called part-of- speech tagging is for the text in word labeling part of speech. Is mainly refers to the concurrently kind of word the lexical category to determine that the concurrently kind of word's lexical category determines only the sentence in according to the context.

B semantic labeling

semantics labeled a word to be equivocal, has formed the word different meanings phenomenon, semantic labeling mainly solves the word different meanings problem. A word equivocal is also in the natural language common phenomenon, but, in certain context, a word can only explain that generally is one semantics.

semantics labeling is to appears the words and expressions semantics carries on the determination in certain context, determined that its correct semantics and labels.

□ Semantic automatic labeling method Usual word is composed of meaning

□ The related word's method which appears using the retrieval context in determines the polysemant right eousnessitem

□ Determines the polysemant using the context matching relations the word meaning

□ To dispel equivocally with the most greatly possible right eousnessitem

C labeling technologies

The commonly used labeling technology route is based on the probability statistics and based on the rule method.

Ⅰ)Based on probability statistics CLAWS algorithm

CLAWS is English Constituent-Likelihood Automatic Word-tagging System (ingredient likelihood automatic lexical category automatic labeling system) one algorithm which the abbreviation, it was in 1983 Ma Shaer (Mashall) when gives the LOB corpus (to have each literary style British English corpus, storage capacity quantity is 1,000,000 words) made automatic part-of-speech tagging proposed

Ⅱ) Based on probability statistics VOLSUNGA algorithm

The VOLSUNGA algorithm is to the CLAWS algorithm improvement, in optimal path's choice aspect, is not only then calculates the probability to accumulate the biggest mark string finally, but along direction from left to right, the use “fortifies at every step” the strategy, regarding the current consideration's word, only retains leads to this word the optimal path, discards other ways, then embarks again from this word, carries on the match this way with next word's all marks, continues to discover the best way, discards other ways, goes forward like this gradually, walks until the entire cross section, obtains the entire cross section the optimal path to take the result output. Counts each word according to the corpus the relative labeling probability (Relative Tag Probability), and is auxiliary the optimal path with this kind of relative labeling probability the choice. The VOLSUNGA algorithm reduced the CLAWS algorithm time order of complexity and the spatial order of complexity greatly, raised the automatic part-of- speech tagging rate of accuracy.

The CLAWS algorithm and the VOLSUNGA algorithm are based on the statistical automatic labeling method, acts according to merely with the present probability labels the lexical category. But, with the present probability is only the biggest possibility and is not the only possibility, by determines the concurrently kind of word with the present probability, is by discards with the present probability low possible premise. In order to enhance the automatic part-of- speech tagging the accuracy, but must auxiliary by based on the rule method, determines the concurrently kind of word according to the language rule.

D other text retrieval labeling technology

Ⅰ) Inverted index

Inverted index is an index structure that contains two hash tables index table, or two B +-tree index table, shown in Table 1, Table 2.

Table 1 Document Table (document_table)

doc_ID	posting_list
Doc_1	t_1_1, ..., t_1_n
Doc_2	t_2_1, ..., t_2_n
⋮	⋮
Doc_n	t_n_1, ..., t_n_n

Table 2 vocabulary (term_table)

term_ID	posting_list
Term_1	doc_1, ..., doc_i
Term_2	doc_1, ..., doc_ j
⋮	⋮
Term_n	doc_1, ..., doc_n

Table 1 is composed of a group of documents record, posting_ list is appears in the documents the word tabulation; Table 2 are composed of a group of word record, posting list is contains this word the documents marking tabulation. Through such two tables, may discover with and assigns the documents related all words as well as with group of word related all documents for the decisive remark collection related all documents. But cannot process the synonym and the polysemant question, and posting list is long, causes the memory expenses to increase.

Ⅱ) Signature File

Features file is a storage database, the characteristics of each record of a document file. A feature of each bit corresponds to a fixed-length string, a bit corresponds to a word, if a particular word corresponds to appear in the document is, then the location of one, otherwise set 0.

Ⅲ. TEXT EXCAVATION PROCESS

The text excavation object usually is group of HTML perhaps the XML form documents collection. Text excavation's general treating processes is: Document Set, Characteristics of the establishment of, Reduced feature set, Learning and knowledge extraction model,

Model Quality Evaluation-Knowledge model.

A . Text Features

Text feature refers to the metadata on the text. It can be divided into descriptive features (text, name, date, size, type, etc.) and semantic features (text, author, title, organization, content, etc.). Text feature to feature vectors, said:,

Where t I for the entry , w i (d) for t iin d in the weights. As the feature vector dimension is usually very high, generally use the evaluation function to carryout feature selection. Evaluation of commonly used functions: information Gain, Expected Cross Entropy, Mutual Information, the Weight of Evidence for Text, Word Frequency.

Document Modeling： Using vector space model(VSM) of the text document model.

Frequency Matrix: line corresponds to the word w, the column vector corresponding to the document d, the simplest vector of values of words in the document appears on the value of 1, otherwise value is 0, Table 3 is based on occurrences of the word for the word frequency vector matrix, the value to reflect the word w and a document d of the correlation.

Table 3 Frequency of the Frequency Matrix document

	d_1	d_2	d_3	d_4	d_5	d_6
w_1	322	85	35	69	15	320
w_2	361	90	76	57	13	370
w_3	25	33	160	48	221	26
w_4	30	140	70	201	16	35

With the similarity of the document word frequency matrix can be measured, the typical method is the cosine similarity metric calculation (Cosine Measure).

□ Calculates frequency matrix the singular value decomposition. Decomposes frequency matrix to become 3 matrix U, S, V. U and V is the orthogonal matrix (UTU=I), S is the singular value diagonal matrix(K×K).

□ Regarding each documents d, after removing in SVD eliminates the word new vector replace original vector.

□ Preserves all vector set, founds the index with the high-level multi-dimensional index technology for it.

□ Carries on the similarity computation after the transformation documents vector.

C studies and knowledge pattern extractions

Ⅰ) Participle： The participle refers to between the text word and the word adds on the blank space, refers to Chinese text, because between English itself word and the word is differentiates by the blank space.

Ⅱ) Automatic participle： The automatic participle is refers to uses the computer adds on the blank space

automatically between the word and the word. The use is:

a) Chinese text automatic retrieval, filtration.

b) Classification and abstract.

c) Chinese text automatic proofreading.

d) Outside Chinese machine translation.

e) Chinese character recognition.

f) Chinese speech synthesis.

g) Take sentence as unit's Chinese character keyboard entry.

h) Chinese character Jan traditional form transformation.

Ⅲ) Main participile method

a) Biggest match law (Maximum Matching method,

Cosine similarity definition: SIM(v,v)□

v1•v2

, is two

MM law): The selection contains 6-8 Chinese characters the strings to take the biggest string, matches in the biggest

1 2string and the dictionary word clause, if cannot match,

documents vectors, The inner product is the standard vector slices off a Chinese character to continue to match, until

found the corresponding word in the dictionary. The match

dot product, Defines for w

i□1 1i 2i

for

B characteristic collection deflation

， is defined direction is from right to left.

1 b) Reversion biggest match law (Reverse Maximum

method, RMM law): The match direction and the MM law are opposite, is from left to right. The experiment indicated: Regarding Chinese, the reversion biggest match law is more effective than the biggest match law.

c) Bilateral matching law (Bi-direction Matching

Term frequency matrix similarly Gao Weishu, sparse data influence, to overcome these questions, the people proposed the latent semantic index (Latent Semantic Indexing) the method reduces the characteristic collection.

Ⅰ) Latent semantic index ： “The singular value decomposes (Singular Value Decomposition using the

matrix theory, SVD)” the technology, transforms the term frequency matrix as the singular matrix (K×K), concrete step:

□ Establishment term frequency matrix, frequency matrix.

method, BM law): Compared with the MM law and RMM law participle result, thus decides the correct participle.

d) Optimum matching law (Optimum Matching method, OM law): The dictionary in word according to them in the text appearance frequency's size arrangement, the high frequency's word arranges before, the frequency low word arranges, thus enhancement match speed.

e) Association backtracking: Uses the mechanism which associates and recalls to carry on the match.

Ⅳ) Feature extraction

The feature extraction is the glossary which, the phrase feature extraction appears to the text.

Characteristic weight function:

Ⅳ. TEXT CLASSIFICATIONS

The text classification is refers to according to the subject category which defines in advance, determines a w(ti)□

(ti)log(1□f

v(ti))□

category for documents set's in each documents. Thus, not only the user can glance over the documents conveniently, moreover may make the documents through the limit

And: Expresses the characteristic weight function; Expresses the characteristic item in the text frequency; Expressed that the characteristic paragraph frequency, namely contains ti the paragraph number/text total paragraph number.

Ⅴ)Automatic digest

The automatic digest is uses the computer to withdraw automatically from the primitive documents reflects this documents center content accurately comprehensively the simple coherent short written work. Our country in 1995 carried on to the automatic digest system has evaluated, the system which for the first time participated has 3. The evaluation result performance is:

1. Three systems may according to the ratio which assigns from the original text select part of sentences.

2. The extraction digest is in the original text sentence, only then in the system 2 digests has rejected some digit.

3. Three system's digests do not superpose nearly completely.

may see the automatic digest system from above result also to have many foundation work to do. the text abstract is refers to from the documents extracts the key information, carries on the abstract or the explanation with the succinct form to the documents.

Thus, the user does not need to glance over the full text to be possible to understand the documents or the documents set overall content. The text abstract is very useful in some situations, for example, search engine when to user returns inquiry result, usually needs to give the documents the abstract.

D. model quality appraisal

Carries on the excavation in the text to be possible to regard as is one kind of machine learning process. The study result is the knowledge model, carries on the appraisal to the knowledge model is the machine learning important component. The typical assessment method is to the text retrieval basic measure.

{relevant}: With some inquiry related documents set.

{retrieved}: The system retrieves documents set.

{relevant} ∩ {retrieved}: Both are related and the actual documents set which retrieves.

precision: Both are related and the actual documents which retrieves with the documents percentage which retrieves.

recall: Both are related and actual documents which and the inquiry related documents percentage retrieves.

hunting zone the search to be easier. At present some websites use the man-power to carry on the classification to the Web documents, some websites use the automatic sorting. The text classification technology algorithm has many kinds, the commonly used algorithm has TFIDF and Nave Bayes and so on.

A Generally method

□ Will have classified in advance the documents take the training regulations.

□ Obtains the disaggregated model from the training regulations (to need test procedure, unceasing refinement).

□ With the disaggregated model which derives to other documents classifies.

B Based on connection taxonomic approach

□ Proposes the key words and the glossary through the information retrieval technology and the connection parsing technique.

□ Uses the existing part of speech production key words and the word concept level (documents category).

□ Discovers the associated word using the connection excavation method, then differentiates each kind of documents (each kind of documents to correspond a group of connection rule).

□ Goes with the connection rule to the new documents classification.

C Web documents automatic sorting

Uses in the ultra link the information to carry on the classification, the commonly used method includes:

□ Statistical method

□ Markov random field (Markov Random Field, MRF)

□ Unifies loose marking (Relaxation Labeling, RL)

Ⅴ. TEXT CLUSTERS

The text cluster and the classified difference lies, the cluster has not defined the good subject category in advance, its goal is divides into the documents certain kinds, the request identical kind in documents content similarity is as far as possible big, but the different kind of between similarity is as far as possible small. Hearst et al. the research had already proven “the cluster supposition” the question, namely approaches with the inquiry related documents cluster's comparison, and is far away from the non-correlated documents. Therefore, the documents which will search using the cluster technology divides into certain

kinds. At present has many kinds of text cluster algorithm. Divides into two big types approximately: Level cluster and plane allocation method.

A Level cluster law

Concrete process:

□ Documents collection D= {d1,⋯, di,⋯, dn} each documents di regards as has single member's kind of ci={di}, these kinds constituted D cluster C=

{c1,…, ci,…, cn};

□ Calculates in C every time to the kind (ci, cj) between similarity SIM (ci,cj);

□ The selection has the biggest similarity kind to arg max SIM (ci, cj), and ci and the cj merge is one new kind ck=ci ∪ cj, thus constitutes D new kind of C= {c1,⋯,cn-1};

□ Is redundant the above step, is only left over one kind until C.

Materially this process constructed one to contain in the kind of level information as well as during all kinds and the kind of similarity spanning tree. Because each time merges time, needs overall situation quite all kind of between the similarity, then choice best two kinds, therefore the operating efficiency is not high, does not suit in the massive documents set.

B Plane allocation method

The plane allocation method is documents collection D=

{d1,…, di,…, dn} horizontal divides for certain kinds, concrete process:

□ The determination must produce kind of number k;

□ Produces k cluster center according to some kind of principle to take the cluster seed S= {s1,⋯, sj,⋯,sk};

□ To D each documents di, calculates it and each seed sj similarity SIM in turn (di,sj);

□ The selection has biggest similarity seed arg max SIM (di, sj), belongs to di take sj as cluster center kind of Cj, thus obtains D cluster C= {c1,⋯,ck};

□ The redundant step 2~4 certain times, by obtains the stable cluster result.

This method speed is quick, but k must determine in advance, seed selection difficulty.

the text cluster also has the k-means algorithm, the simple Baye cluster law, the K- most close neighbor to refer to the cluster law, the graduation cluster law as well as based on the concept text cluster and soon.

Ⅵ. RELATED CONTENTS

Text excavation besides above several introduction content, but also has the following related content research:

□ Chinese character input and Chinese corpus.

□ Text phrase delimitation and syntax labeling.

□ Electronic dictionary construction.

□ Terminology database.

□ Machine translation.

□ Computer auxiliary text proof reading.

□ Information automatic retrieval system.

□ Chinese speech recognition system.

□ Chinese speech synthesis system.

□ Chinese character recognition system.

The related text excavation's product model has the IBM text intelligence excavator (the hard core is Text Miner, its major function is the feature extraction, the documents accumulation, the documents classification and the retrieval; Supports 16 languages many kinds of form text data retrieval; Uses the deep level the text analysis and the index method; Supports the full-text search and the index search, the search condition may be the natural language and the Boolean logical condition.), the Autonomy Corporation most core's product is Concept Agents (can extract concept automatically from text) as well as T singhua University's TH-OCR Chinese character recognition system (recognition precision reaches above 98%).

Ⅶ. CONCLLUSIONS AND FORECAST

The text excavation, needs to use the natural language processing technology inevitably, constructs the large-scale real text the corpus is the most foundation work. This article elaborated the content is in the text excavation key job. if the foundation work is not solid, the text excavation is very difficult on a big stair. Basic research's foresightedness ought to be able to guarantee in technical the sophistication. Future text excavation technology should be the knowledge retrieval, the knowledge retrieval development should be able the effective addressing following some key questions: (a). Structurized data and non-structurized data mix retrieval; (b) Half structurized content retrieval XML content retrieval; (c).Engine intellectualization knowledge retrieval.

REFERENCE

[1] C.Faloutsos. Access Methods for Text. ACM Computer. Survey. , 17 p49-74, 1985.

[2] G.Salton. Automatic Text Processing. Reading ,MA: Addison-Wesley,1989.

[3] C.J.Van Rijsbergen. Information Retrieval.
Butterworth,1990.

[4] C.T.Yu and W.Meng. Principles of Database Query Processing for Advanced Applications. San Francisco: Morgan Kaufmann,1997.

[5] K.Wang,S.Zhou,and S.C.liew. Building Hierarchical Classifiers Using Class Proximity. In Proc.1999 Int. Conf.VLDB’99,P363-374,Edinburgh,UK,Sept.1999.

[6] P.Raghavan. Information Retrieval Algorithms: A Survey. In proc.1997 ACM-SIAM Symp. Discrete Algorithms, p11-18,NewOrleans,LA,1997.

[7] J.M.Kleinberg and A.Tomkins. Application of Linear Algebra in Information Retrieval and Hypertext Analysis. In proc. 18th ACM Symp. Principles of Database Systems,P185-193,Philadelpgia,PA,May1999