'Research Journey' International Multidisciplinary E- Research Journal
Special Issue I – Emerging Trends & Issues in Research & Development
Global Impact Factor – 0.565 (2014) & 0.676 (2015)
ISSN : 2348-7143 (Online)
April-May-June - 2016

AUTO LABELING OF DOCUMENT USING CLUSTERING TECHNIQUE

Mrs. Smita P. Metkar
M.S.G college Malegaon camp
Smitapmetkar27@gmail.com

Abstract :

Automatically labeling of document with words prepared in clusters which show their concept is tedious to implement. Inspiring from the model that forms words like a hierarchy of different elements of different classification. The author in this paper work representing the methods for document labeling using concept of clustering. In one method, author considering the existence of a document hierarchy labeling with the most guess words. Whereas in another method based on selected words which both occur frequently in a cluster and effectively discriminate the given cluster from the other clusters and their labels produced and all this is presented with superior results to the commonly employed methods.

Keywords: *labeling, document hierarchy, clustering, prediction, frequent, dependencies.*

Introduction

In the world of internet for retrieving, surfing and managing the on line data there is numerous techniques of research. For that purpose we cluster the document into relevant groups. Mostly the document clusters given hierarchy. Whenever user browse user would quickly scan the cluster in which they are interested. user of clusters does not goes in the detail. e.g. for browsing large document collections Scatter/ Gather technique is used [1]. After eliminating the stop words, for e.g. *for, the, and* [1, 2] clusters are usually labeled by their most repeated words. Subject is mostly shown by these repeated words whereas it cannot show the cluster details as they are reduces with the collection specific stop words. e.g suppose there is a collection of computer science research paper the words which are commonly use are *paper, method, result, system* are very usual to computer science controls hence there is no need to give more information to somebody who knows that documents are of research papers of computer science. Someone can also use the guess words for the given clusters. That guess word of the cluster are also known as the silent words or key words which are unambiguation[3]. Sometimes titles of the papers which are more near to cluster [1] are also labeled to clusters it is more suitable and personal so fail to give judgment into the quality of clusters under application.

The author introduces various ways of selecting the words for labeling of cluster that helps to avoid the prior problems. In one method estimate the presence of document hierarchy, which is

constructed manually or populated or grouping resulting from use of different hierarchical clustering algorithm. In second way we select those words which have both properties means occur often in a cluster and completely separate the given cluster from the other clusters.

Methods For Document Labeling

X^2 Method : When count of data is available the X^2 Method is more useful . Working of this method is when each word at each node in a group start from the root node and again recursively moving lower to the group. the hypothesis cannot be rejected when a word is like to come at every child of a given node. When it is occur it is marked as general to that subtree and the changes to the current node's sequence of the node. Working of this method is shown by this algorithm as follows:

Input: A group of documents where child contains words from all parts of the documents and all leaf union together.

1) Starting from the leaves and goes to the root occupy all the intermediate node by combining the bags of words in its children.

2) For discovering the dependencies the X^2 test starts for the root and do for each word

a) We conclude that probability of occurring a word in a children is different when this test deny the independence assumption which is more specific for one or more categories down the tree;

b) We conclude the probability of occurring word is equally likely in all of the children when this test not pass to accept the independence hypothesis. Keep the word at the current node and declare it as general at the working node to the subtree. Delete all such type of words from every nodes under the node where this test was perform;

3. Repeat step 2 again and again and moves down the tree to the leaves.

Output: Starting hierarchy of the given document and the hierarchy of the words is same and at each node there are specific words which are not available in the subtree under the current node. Relating to cluster of the given documents which we want to label a label is the list of often occurring words.

Predictive and Frequent Words Method

In this method words selected for labeling are depend upon the multiplication of local frequency and productiveness can also be demonstrated by the example as the "frequent and predictive words" method, words are selected as labeling based on.

The product of local frequency and productiveness:

$$(p(\text{word} \mid \text{class}) \times p(\text{word} \mid \text{class})) / p(\text{word})$$

A researcher, Yarowsky[3] in his work used this feature of combined use of local frequency and productiveness. This is used for choosing very vital words in class for demonstrate his view of word sense which are not ambiguous. By all this we also know that this method is not applicable to label document clusters. In the given mathematical expression it clears that there are two parts which have well defined meaning i.e. the first part is given for the productiveness p(word | class) / p(word) is near to share information estimator[4] and TF−IDF [5] count use in information fetching. In this scheme the author gives more weight to that words which occur often in a given cluster and less weight is given to those words which occur after in all the clusters; **p(word | class)** is count of the word in a given cluster and **(word)** is the word's count in a more general category or in the total collection. The words which get high weight are good classifier for differentiating one cluster form another. By using this formula we choose two types of words which occur frequently in the cluster and very particular to the cluster. In next part the author shows some experimental results and also illustrate the above methods by comparison.

Experimental Results

The core search service [6] select the document for experiment from the abstract of the various research papers of computer science which are related to AI and Information retrieval was used. See diagram -1

Diagram -1 Cora Hierarchy Subset

Root

Artificial Intelligence

- Agents
- Expert Systems
- Data Mining
- Knowledge Representation
- Machine Learning
- Games and search
  - Case Base
  - Genetic Algorithm
  - Probabilistic Methods
  - Rainforcement Learning
  - Theory
  - Rule
- Natural Language Processing

Planning

Speech

Robotics

Vision and pattern recognition

Information Retrieval

Extraction

Digital Libraries

Retrieval

Filtering

The research papers presented by Cora where collect together and stemmed using Porter stemmer [7] in which the stop words and the words which are not occur five or less time are deleted. This experiment results in 1022 specific words. This count is not so large data set so it gives feasible presentation purpose well.

For low count data the X^2 method is unreliable so we applying it whenever the word counts are more than five that is a frequently used rule for this method. At the root node the X^2 test were execute at its particular level of 0.95. The author attempt this test in 122 case and he did not detect the dependancies in children. After testing these 122 cases these words are keep at the root as general to all the documents collection. Following are the most frequent words :

method, result, present, paper, approach, data, set, new, task, techniqu, agent, domain, search, differ, number, application, work , knowledge, system, develop.

Above is the list of words which are more common to all research papers of computer science

Table 1 shows the some highest rank word labels of the most commonly used and most predictive words it is compared to both method means first most frequent words and second most predictive words it is for the information retrieval node which may be a interior node.

Table 1. Information Retrieval

Most frequent	Most frequent	Most frequent and predictive
Inform	area	docum
Text	attribut	inform
Us	alloc	text
Docum	compute	retriev
Result	io	librari
Learn	effici	web

The word use (referred as us) present very often in all the collection of documents and weighted by the most frequent and predictive words method. In case of information retrieval web is a very commonly used word but added in the locally frequent words it is added in most predictive and frequent words due to its up weight by this algorithm is applied to it. Some of the examples of the labels which are given to the leaf node are given in the following tables which are titled by the cora cluster :

Table 2. Natual Language Processing

Most frequent	Most frequent	X^2	Most frequent and predictive
Us	bank	us	tag
present	symbol	tag	word
approach	grammar	algorithm	pars
Word	lexic	inform	corpu

Table 3. Robotics

Most frequent	Most frequent	X^2	Most frequent and predictive
Task	mobile	behaviour	map
behaviou	obstacl	control	environ
Plan	multirobot	research	manipul
Learn	sensori	perform	navig

Table 4. Agents

Most frequent	Most frequent	X^2	Most frequent and predictive
Agent	action	inform	agent
model	present	research	negoti
Inform	call	us	distribut
Paper	capabl	distribut	Coordin
distribut	fac	environ	team
Us	distribute	model	multiag

Here one thing is noted that agent is not the part of the X^2 method's list for the "Agents" cluster in table 4 this is unsatisfactory but occur due to agent was used in information retrieval and artificial intelligence so it is keep at the root.

Conclusion and future work

In this method the author gives various methods of labeling document by using cluster and choosing topic relating keywords. In most predictive and frequent method label produced are the

best labels. Gathering all the words which occur frequently in the given clusters and strongly representing the cluster from the other clusters. X^2 method method also does the performance on most predictive and most frequent words but this method is very sucessful in the identification of *document specific stop words*. Unfortunately, none of the method gave a satisfactory result for the internal nodes of the group. This shows the well feature of document collection, representaion that the methods realating to the internal nodes , information retrieval, Artificial intelligence and machine learning are very diverse which uses and contain large no of topics.

References

[1] Cutting, D. R., Karger, D. R., Pedersen, J. O., and Tukey, J. W., Scatter/Gather, "a Cluster−Based Approach to Browsing Large Document Collections", 15th International ACM SIGIR Conference on Research and Development in Information Retrieval, 1992

[2] Sahami, M., Hearst, M., and Saund, E., Applying the Multiple Cause Mixture Model to Text Categorization, In ICML−96: Proceedings of the Thirteenth International Conference on Machine Learning, pp. 435−443, San Francisco, CA: Morgan Kaufmann,1996.

[3] Yarowsky D.,"Word−Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora ,"In Proceedings, COLING−92. Nantes, 1992.

[4] Cover, T. M., and Thomas, J. A., Elements of Information Theory, New York

[5] Salton, G., Developments in Automatic Text Retrieval, Science 253, Aug. 1991.

[6] Cora, Computer Science Research Paper Search Engine, http://www.cora.jprc.com/

[7] Porter, M. F., An Algorithm for Suffix Stripping, Program, 1980.