Clustering Engine for Desktop Usability

Prajakta Sharmao Pawar (ME-IT), Sushopti Gawade (guide)

Abstract: *The increased capacity and lower prices of computer hard drives led to a new universe of knowledge. Although search and information retrieval techniques are already widely used in the Internet, its application in personal computers is still incipient. The management of information is relatively difficult when it comes to large size of data lying in the machine. There is a need to manage the number of documents and the data retrieval is difficult when the number of files is too many.*

This project proposes an approach of document clustering as user need to retrieve information from set of documents.

[1]Document clustering is one of the approach in clustering method which is specifically used for management and retrieval of documents.

[2]The clustering engine intelligently works towards providing user with document as input, the specific documents with output.

I. Introduction

Enterprise data mining applications often involve complex data such as multiple large heterogeneous data sources. In such situations, a single method or one-step mining is often limited in discovering informative knowledge. It would also be very time and space consuming, to relate relevant large data sources for mining patterns that consist of multiple aspects of information. It is crucial to develop effective approaches for mining patterns combining necessary information from multiple relevant business lines, catering to support decision-making actions rather than just plain retrieval. The recent years have seen increasing efforts on mining more informative patterns, e.g., integrating frequent pattern mining with classifications to generate frequent pattern-based classifiers. [1][2]

This project builds on our existing works and proposes combined mining as a general approach to mining for informative patterns combining components from either multiple data sets or multiple features or by multiple methods on demand. This project tries to summarize general frameworks, paradigms, and basic processes for multi-feature combined mining, multisource combined mining, and multi-method combined mining. Novel types of combined patterns, such as incremental cluster patterns, can result from such frameworks, which can be directly produced by this method.

II. Literature Survey

Paper[1] basically explains the existing clustering engines such as Carrot2, Google Desktop, Aduna AutoFocus. It explains the problems in the existing web clustering engine and slowly takes us towards the desktop clustering approach of ICE- Intelligent Clustering Engine. It puts focus on the difference between existing clustering engines and the proposed ICE system. From this paper, the idea of document clustering for desktop has been derived. This idea would help to develop a clustering engine for desktop which a completely new application for document clustering. Paper [2] explains various aspects of web clustering engines. It explains the entire process of Web Clustering Engine. It also specifies the components of web clustering engine and explains each component in details. This paper guides the detailed concepts of web clustering engine. Web clustering engine is a base for intelligent clustering engine and it also gives an overview of the basic components of web clustering engine. Paper [3] focuses mainly on Carrot2 clustering Framework. It explains the Carrot2 framework with proper diagram. A complete overview of Carrot2 is provided in this paper. This paper guides by providing details of carrot2 clustering engine. It aids the project by providing the carrot2 clustering engine as a base for intelligent clustering engine. Carrot2 clustering engine is based on web search engine and desktop clustering engine is based on local desktop documents data. Paper [4] explains the approach to web clustering engines and give information regarding the survey of the clustering engines. It also includes advantages of Web Clustering Engines over the existing search engines. The drawbacks of search engines are given in this paper. The survey provided in this paper gives the right information that which approaches and base should be used to develop and implement the Intelligent Clustering Engine. Paper [5] gives information about information architectures. It also explains the approach in Carrot2 framework. It discusses the process used in Carrot2 framework. It includes the details of how the final output is obtained in the Carrot2 engine.

III. Algorithms Of Intelligent Clustering Engine

Creating a similarity matrix of snippets and assigning weights to documents
Algorithm1: Term-weighted Similarity Measure [9]

Steps

1. Let () n Rx be the set of top n documents returned by a search engine when using x as the query.
2. For each document in d R x ∈ , construct the term vector iv with TF×IDF and truncate each vector i v to include its m highest weighted terms.
3. Let C(x) be the center of the normalized vectors v_i :

$$C(x) = \frac{1}{n}\sum_{i=1}^{n} v_i / \|v_i\|$$

4. For each element $k_i \in C(x)$, suppose i t is the term corresponding to k_i . Construct the term vector () WC x with $k_i \times \frac{E(t_i)}{\sqrt{D(t_i)}}$, where $E(t_i)$ and $\sqrt{D(t_i)}$ are the expectation and unbiased deviations of term frequency of it respectively.
5. Let QE(x) be the normalization of the center () WC x :

$$QE(x) = \frac{WC(x)}{\|WC(x)\|}$$

So, the similarity of two short snippets can be obtained by calculating the inner-product of
QE(x)×QE(y).

n- number of documents
Rx-Query
x- xth document
dR- R documents in sample
TF- Total frequency
IDF- Inverse document frequency
Cx- Center of vectors
Vi- Vector i
Ki- element i
E(t) Expected Deviation of term frequency
C(x) - C(x) be the center of the normalized vectors v_i
QE(x)- QE(x) be the normalization of the center () WC x
WC(x) weighted center of term

IV. Module Description Of Ice

Loading Dataset

This module will load the dataset into the software. The dataset is a collection of a set of documents.

Search Keyword

This module will take the keywords which the user is interested in. The keywords will be taken as input from the user and will be searched individually.

Preprocessing

This module will perform preprocessing upon the search keywords. Each search keyword will be analyzed in detail. Every form of individual keyword will be considered including the synonym of each.

Calculating Frequency

A particular search keyword will exist in a document multiple times. We will calculate the frequency of each keyword..

Doc1-	48
Doc2-	40
Doc3-	12
Doc4-	33
Doc5-	9

In Groups-

- Tables
- Graphs
- Reports
- Maps

V. Conclusion

Clustered Grouping the data in a systematic manner such that we get solid dimensions with this research engine is been concluded which is given by example Bing. The growth year of research use data had been extracted by [4][1] [3] [3] [2] [2] to papers [3] [2] [1] [2].

Reference

[1]. Intelligent Clustering Engine: A clustering gadget for Google Desktop, 2012 Lando M. di Carlantonio a,⇑, Bruno A. Osiek a, Geraldo B. Xerox a, Rosa Maria E.M. da Costa

[2]. Web Clustering Engine based on Wikipedia, Yuvarani Meiyappan1, N. Ch. S. Narayana Iyengar2 and A. Kannan3 1Lead, Infosys Limited, Bangalore, Karnataka, India 560100

[3]. A Survey of Web Clustering Engines CLAUDIO CARPINETO, STANISlAW OSIN´ SKI-2012

[4]. Carrot2 s.c.t. Clustering Framework, Dawid Weiss, Pozna´n University of Technology, Pozna´n, Poland and Stanisław Osi´nski

[5]. A Document Clustering Approach for Search Engines Chun-Wei Tsai, Ting-Wen Liang, Jiun-Huei Ho, Chu-Sing Yang and Ming-Chao Chiang

[6]. An Efficient Algorithm for Clustering Search Engine Results -- Hui Zhang, Bin Pang, Ke Xie, and Hui Wu National Laboratory of Software Development Environment Beihang University Beijing 100083, China {hzhang, pangbin, xieke, wuhui}(nlsde.buaa.edu.cn

[7]. Link Based K-Means Clustering Algorithm for Information Retrieval M.Sathya#, J.Jayanthi*, N. Basker# #Department of Computer Science and Engineering Sona College of Technology, Salem, TN, India

[8]. Web Document Clustering Using Document Index Graph B. F. Momin 1 Student Member IEEE, P. J. Kulkarni 2, Amol Chaudhari 3 1bfmomin@rediffmail.com, 2 pjk_walchand@rediffmail.com

[9]. An improved measuring similarity for short text snippets and its application in Clustering Search Engine, Zhao Li, Hong Peng, Peng Peng, Xi-Ping Jia, Jia-Bing Wan

[10]. Website1:Aduna–AutoFocus.<http://www.aduna-software.com/technologies/autofocus/overview.view> Visited July 2009.Carrot Search (2011a).

[11]. Website2:Carrot2 clustering engine. <http://search.carrot2.org/stable/search> Visited September 2011.