Efficient Term Frequency and Optimal Similarity Measure of Snippet for Web Search Results

D.Rohini [#1], R.Janaki [*2]

[1] *M.Phil. Scholor, Department of computer Science, Mother Teresa Women's University, Chennai – 15*
dv.rohi@gmail.com

[2] *Assistant Professor, Department of computer Science, Queen Mary's College (A), Chennai – 04*
janakigsr@gmail.com

Abstract— All clustering methods have to assume some cluster relationship among the data objects that they are applied on. Similarity between a pair of objects can be defined either explicitly or implicitly. In this paper, we introduce a novel multi-viewpoint based similarity measure and two related clustering methods. The major difference between a traditional similarity measure and ours is that the former uses only a multi-viewpoint on clustered, which is the origin, while the latter utilizes many different viewpoints, which are objects, assumed to not be in the same cluster with the two objects being measured. Using multiple viewpoints, more informative assessment of similarity could be achieved. It combines the neighbourhood preservation capability of multidimensional content with the familiar optimal snippet-based representation by employing a multidimensional content to derive two-dimensional layouts of the query search results that preserve text similarity relations, or neighbour hoods. Theoretical analysis and empirical study are conducted to support this claim. Two criterion functions for document clustering are proposed based on this new measure. We compare them with several well-known clustering algorithms that use other popular similarity measures on various document collections to verify the advantages of our proposal.

Keywords— Multi-view point, term frequency (TF), clustering, Euclidean distance

1. Introduction

Users searching for information are faced with the challenge of how to explore the many documents retrieved by a search engine. Users expect that the results are displayed in a way to make it easy for them to identify the documents that are most likely to be relevant. Commonly, search results are presented as a ranked list, which has the advantage that users know where to start their search for relevant documents. However, users have to move sequentially though the list and only a small subset of the documents is visible in a single screen. Many visual interfaces have been developed to increase the number of documents that users can explore in a single view (Hearst, 1999; Mann, 2002). These visualizations can be useful when users are not just looking for a few relevant documents, but need to find a greater number of relevant documents or want to gain insight into how a large number of documents are related to their search interest. The exponential increase in the information available on the web requires efficient organization and visualization systems to facilitate faster access to information [4]. A typical application of a visualization system in the context of information retrieval on the web is to visualize the search results of a query launched on a search engine [3].

Search engines such as Google; tend to return a long list of search results with titles, small images and short paragraphs. Users have to open each and every web page to assess its utility and relevance to the searched topic which can become tedious and unproductive. As the information available on the web increases, there is an obvious need to organize and visualize these results [4]. Visualization of clusters in small world and scale free networks remains a challenging problem. Moreover representing web search results such that users can easily understand the content and navigate through the web pages remains an active area of research. In this paper we present a system that combines a previous clustering technique with a novel method to layout the clusters. This approach combines several established methods from the domain of information visualization and graph layouts to present web search results. These algorithms are adapted to cater the needs of underlying data and explicitly assist in its analysis, which in our case would be the different themes revolving around the searched topic. Moreover, the visualization also allows us to visualize the relationships between these themes by representing the words that relate different clusters.

2. Related Work

Different visualization systems for web search results can broadly be grouped into two categories: List Based Systems and Graphical Visualization systems. The list based systems keep the traditional ordered list visualization adding visual aids such as bolding words in the paragraphs [5] or clustering web pages and presenting a tree view [5][3] along with the list. The web pages are displayed at the bottom screen as we navigate through different searched keywords. The elements are not clustered which makes it difficult for the user to have an idea about the topics revolving around the key words searched.

Given an object browsed by the user, recommendation algorithm generates the recommend list of objects (e.g. similar pages) for the user to choose. Under circumstance of information overload, people usually prefer being recommended to searching. On another hand, recommendation plays key role in long tail mining [2], which include clustering objects into categories, comparing various products and catering various non-mainstream demands [2]. Recommender systems, which have been widely used in e-business, Netnews and online music sharing [3], aim to lead user to find his desired products. The global ranking (GR) methods rank objects using single criteria such as video viewed time. Content based method (CB) methods analyse the correlation between objects.

3. Proposed : Context- contextual Aware Recommendation system

3.1 Context Clustering

Context-awareness allows software applications to use information beyond those directly provided as input by users. This information becomes important in an environment where applications are accessed through mobile and ubiquitous devices that communicate with each other. This work has a general goal of facilitating services selection, exploring the synergy between these two research areas: recommender systems and context- aware computing. Clustering is a fundamental operation used in unsupervised Context collection in a data center, automatic topic extraction and information retrieval. The similarity measure-based CPI method focuses on detecting the intrinsic structure between nearby Context Collections rather than on detecting the intrinsic structure between widely separated Context Collections. Since the intrinsic semantic structure of the Context Collection space is often embedded in the similarities between the Context collections. The low-dimensional representation of the ‘i’th document in the semantic subspace, where i =1, 2, 3….n.

$$\max \sum_{i}^{n} \sum_{xj \in N(xi)} \mathrm{corr}(Ti, Tj)$$

and

$$\min \sum_{i}^{n} \sum_{x!\ j \in N(xi)} \mathrm{corr}(Ti, Tj)$$

D1 = If two documents are close to each other in the original document space, then they tend to be grouped into the same cluster ; D2 = If two documents are far away from each other in the original document space, they tend to be grouped into different clusters. Where N(xi) denotes the set of nearest neighbours of xi. The equivalent metric learning,

$$d(x,y) = \propto * \cos\,(x,y)$$

Where $d(x,y)$ denotes the similarity between the document$(x \text{ and } y)$, α corresponds to whether ‘x’ and ‘y’ are the nearest neighbours of each other. This contextual information is done explicitly or even implicitly, it should be conducted as a part of the overall data collection process all this implies that the decisions of which contextual information should be relevant and collected for an application.

3.1 Context – Contextual pattern analysis

The majority of clustering models, which includes those data or context word/ phrase used in subspace clustering, define similarity among different pages/web link sequence by distances over either all or only a subset of the dimensions. This method of using sentence structure to derive semantic relationships is to define a set of semantic or distributional features, and use this pages user log feature sets for some classified similarity/related estimation. Many different video user log feature sets have been proposed, usually with regard to some notion of the word's context and suited to contextual information. We have user log feature set are identify by the text occurrence of the word/ context in a set of context file where the similarity between two context words was defined by the number of times they co-occurred in the same page context information.

Context-contextual pattern method use co-occurrence of words in the same set of documents to determine the equivalence, but here the distance measure was the mutual information between the words, calculated from these co-occ. Our proposed recommender systems are built based on the knowledge of partial user preferences, i.e., user preferences for some (often limited) set of items, and the input data for traditional recommender systems is typically based on the records of the form < User; item; context; video>. In contrast, context-aware recommender systems are built based on the knowledge of partial contextual user preferences and typically deal with data records of the form < user; item; context content, webpage.

Where each specific record includes not only how much a given user liked a specific item, but also the contextual information in which the item was consumed by this user.

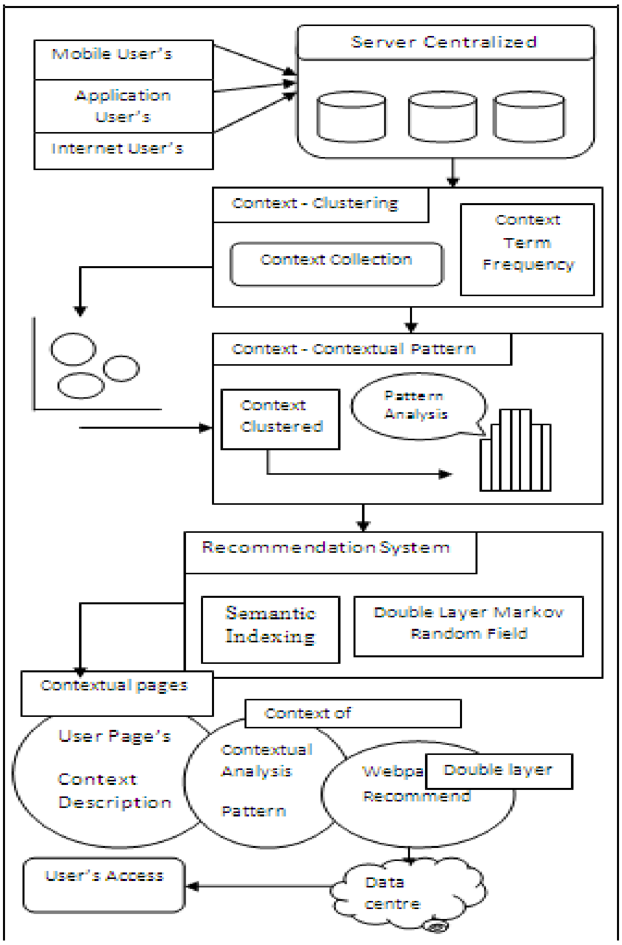

Fig. 1: Architecture of term frequency and similarity measures

3.3 Recommendation system based on Context relational Markov Random field (CrMRF)

Context is a difficult concept to capture and describe; fuzzy ontology's and semantic reasoning are used to augment and enrich the description of context. Similar [A,B] include two kinds of similarity, Similar Sem(A,B) denotes semantic similarity of keywords, SimAttr(A,B) represents similarity of attributes, minSD(ai,B) is minimal semantic distance between keyword 'ai' and all keywords in B, and minSD(bj,A) is minimal semantic distance between keyword 'bj' and all keywords in A. The semantic distance is calculated according to semantic tree/dictionary. Markov Random Fields (MRF's), also known as markov random networks are a common way to model the joint probability of a group of random variables. Such *Context relational* MRFs (CrMRF's) may reprocess the same potential function for many factors in the instantiated model. This means that the model uses *shared parameters* that allow reasoning about a set of variables as a group. Based on our proposed learning on CrMRF from experiential evidence we are given a set of training samples D = {x[1], ..., x[M]}, each is an assignment to the variables X (In this work we focus on the case of fully experiential data, which means that in each sample values are assigned to all the variables in X). Our goal is to learn an appropriate set of features F = {f1... fk} (*Feature Selection data related to video*) and their corresponding parameters θ = {θ1... θk} (*Parameter Estimation*). In otherwise, we want to construct the best generative model for the given evidence.

Lemma1: A Context relational MRF scheme. S is defined by a set of types T, their attributes A and a set of template features $f = \{f1, \ldots fk\}$. A model is a scheme combined with a vector of parameters $\theta = \{\theta i, \ldots .\theta k\} \epsilon\ Rk$…
$P(w : S, T, \theta) = 1/Z(\theta, T) exp \sum_{i=1}^{n} \theta\, i. fi(w)$…

The context source also provides a context warehouse where context information can be stored. Context updater is the component which updates the context information at the context warehouse. The context source later distributes this context information to the subscribed clients. The context source, on registration with a service directory provides a reference to its video's without any redundancy.

Table 1: Cluster–Context Contextual relation

S.No	Clusters Name	#Clusters	Cluster Realtion [Cr]	Cluster without Realtion [Cr]
1	Context-Contextual relation	24	24	21
2	Context Cluster	23.3	22	17
3	Context Information	20	17	13
4	Context-Video Group	22 (**Group content)	20	19
5	Context-Video Profile	21(**Profile content)	19	16

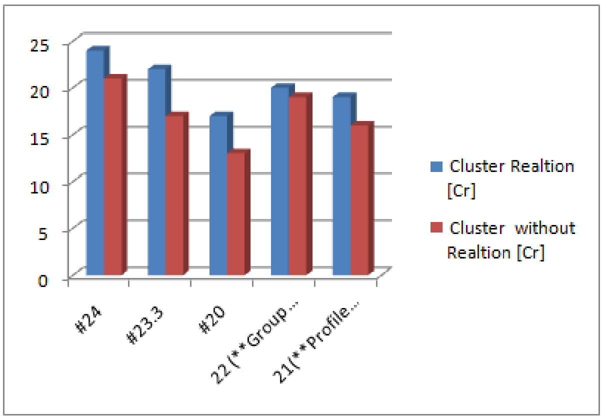

Fig: 2. Graphical representations of Clusters – Context Contextual relation & without relation

4. Conclusion

Recommender systems lead to major progress over the long period of time in multimedia area like YouTube, video file sharing which improve the expected competence of suggestion in video recommender systems. The consistent empirical estimation of the proposed technique provides. We compare the real-time recommendation latency brought by three methods, such as Cluster without relation, rule-based algorithm without well optimization, and Context: Contextual rule-based algorithm with optimization. The result is given in Fig. 3.3. The figure shows that Cluster relation Context: Contextual rule-based algorithms reduce latency about six times rather than CF [3]. If the rule-based algorithm is optimized by graphical representation, have the latency will be reduced for upto 69% and achieved the contextual analysis reduced video redundancy or unrelated video.

References

[1] D. Li, Q. Lv, X. Xie, L. Shang, H. Xia, T. Lu, and N. Gu, "Interest based real-time content recommendation in online social communities," Knowledge.-Based Syst., vol. 28, Apr. 2012.

[2] M.-H. Kuo, L.-C. Chen and C.-W. Liang, "Building and evaluating a location-based service recommendation system with a preference adjustment mechanism,"Exp. Syst.Appl., vol. 36, no. 2, pp. 3543–3554, Mar. 2009.

[3] Z.-D. Zhao and M.-S. Shang, "User-based collaborative-filtering recommendation algorithms on Hadoop," Proc. WKDD, 2010, pp. 478–481.

[4] M. J. Pazzani and D. Billsus, "Content-based recommendation systems," The Adaptive Web. Berlin, Germany: Springer-Verlag, 2007, pp. 325–341.

[5] Z. N. Chan, W. Gaaloul, and S. Tata, "Collaborative filtering technique for web service recommendation based on user-operation combination," Proc. OTM, 2010, pp. 222–239.

[6] Razvan C. unescu, Raymond J. Mooney "Relational Markov Networks for Collective Information Extraction", ICML-2004 Workshop on Statistical Relational Learning and its Connections to Other Fields (SRL-2004).

[7] Hammami, M. ; Chahir, Y. ; Liming Chen "Web Guard: a Web filtering engine combining textual, structural, and visual content-based analysis " Knowledge and Data Engineering, IEEE Transactions on Feb. 2006

[8] Chao Zhou ; Key Lab. of Machine Perception, Peking Univ., Beijing, China ; Yangxi Li ; Bo Geng ; Chao Xu "Learning a Scalable Ranking Model for Content Based Image Retrieval " Multimedia and Expo Workshops (ICMEW), July 2012.