نخستین کنفرانس بین المللی پردازش خط و زبان فارسی
۱۵ و ۱۶ شهریور ۱۳۹۱
دانشگاه سمنان - دانشکده مهندسی برق و کامپیوتر

Semantically Clustering of Persian Words

Alireza Arasteh
Payame Noor University, Qom branch, Qom, Iran
Computer Research Center of Islamic Sciences
Qom, Iran
arastehalireza@gmail.com

Mohammad Hossein Elahimanesh
Islamic Azad University, Qazvin branch, Qazvin, Iran
Computer Research Center of Islamic Sciences
Qom, Iran
elahimanesh@noornet.net

Ahmad Sharif
Islamic Azad University, Science and Research branch,
Tehran, Iran
Computer Research Center of Islamic Sciences
Qom, Iran
asharif@noornet.net

Behrouz Minaei-Bidgoli
Iran University of Science and Technology, Tehran, Iran
Computer Research Center of Islamic Sciences
Qom, Iran
b_minaei@iust.ac.ir

***Abstract*— Clustering is one of data mining task which aims to divides a set of objects into groups so that similar objects fall into the same group and objects with different features are put into different and separate groups. This paper presents a technique for semantic word clustering which is one of the applications of data mining techniques in the task of natural language processing. Word clustering is used in various fields of text mining such as word disambiguation, information retrieval, language modelling, and text classification. This paper proposes a graph based method to clustering Persian words. The proposed method is a type of pattern-based clustering. This method includes two parts; in the first part using statistical similarity measures such as Chi-Square, pointwise mutual information (PMI), and Cosine a word co-occurrence graph is obtained. In the second part, the graph is further divided into appropriate clusters by Newman's graph clustering algorithm. Our researches show that Chi-square is the best measure to cluster the words in Persian.**

Keywords-component; Word Clustering; Text Mining; Persian NLP, Graph-base Clustering.

I. INTRODUCTION

Word clustering is the task of divides a set of words into groups so that words within a group are closely related and have no strong relation with words in other groups. Relationships between and among words can be semantic or derivational. For example words such as Ping-Pong, world cup and football are semantically related with label sport whereas derivational relationships present words with the same root.

Rapid grows of textual data in the world has increasing necessitated the application of data mining techniques. Word clustering is one of the practical techniques which on the one hand can improve performance of the text mining[1] application and on the other hand can reduce the dimensions of textual data.

Some natural language processing applications such as question answering, document clustering and text classifying employ word clustering techniques. For instance Momtazi and Klakow used a type of word clustering technique to the task of Question Answering (QA) (2009). They employed the word clustering technique to increase the efficiency of sentence retrieval in QA systems. Their results showed that using of word clustering can improve the average precision of data retrieval system from %23.62 to% 29.91.

In the previous research, word clustering is also used to reduce dimensions of the data in text classification processes. Baker and McCallum (1998) believe that word clustering is better than other techniques such as Latent Semantic Indexing, Class-based clustering, Feature selection by mutual Information and Markov-blanket-based feature selection to reduce the dimension of textual data.

This paper presents a two part method for clustering words. In the first part of the proposed method, various techniques for making word relationships graph have been examined. In the second part, employment of Newman's graph clustering technique which has yielded good results in the previous studies (Matsuo et al., 2006) is explained.

The rest of the paper has been organized as follows: In section 2 the related literature is reviewed and in section 3, a detailed description of the proposed algorithm is presented. Evaluation methods and results of evaluation are discussed in section 4. Conclusion and future work are stated in the last sections (section 5 and 6).

[1] Text mining is one of data mining subfields.