Question Answering System using PLSA

Anurag Gautam , Harshit Maheshwari
Advisor : Dr. Amitabh Mukherjee
{ganurag, harshitm , amit} @cse.iitk.ac.in
Department of Computer Science and Engineering ,
IIT Kanpur , India
Feburary 20 , 2012

1 Introduction

In order for the computers to interact with the users more naturally the computer must understand and automatically infer what the user wants to say from what he actually says . Question Answering System (referred to as QAS henceforth) is basically a search program that strives to give best possible answers to the questions asked from it by the users from the database it has been provided. Probabilistic latent semantic analysis (PLSA), also known as probabilistic latent semantic indexing (PLSI, especially in information retrieval circles) is a statistical technique for the analysis of two-mode and co-occurrence data. PLSA evolved from latent semantic analysis, adding a sounder probabilistic model. PLSA has applications in information retrieval and filtering, natural language processing, machine learning from text, and related areas.

2 Usefullness

The basic principle behing the question answering system is that the users do not want to read the entire documents ,especially when they have a specific query. In this project we aim to acheive this by catering to the specific requests by the users and generating the most accurate answer possible to the query. Such question answering system is therefore also useful for the search engines. Though due to the multicontexual nature of english language this task becomes difficult but still the PLSA algorithm aims to capture the intended meaning from the queries and match it to the latent topics of the documents .

3 PLSA (Probabilistic Latent Semantic Analysis)

Given a document S, a term set W and a topic set Z , the conditional probabilty of sentence-term P(s,w) can be described as follows :

$$P(s,w) = P(s)\sum_{z \in Z} P(w|z)P(z|s)$$

where P(w|z) is the conditional probability of words in lantent semantic topics , P(z|s) is the conditional probability of topics in the documents. This is in accordance with the EM algorithm and exports the optimal P(Z) , P(W|Z) and P(Z|S) . The query built from the question is mapped into the topic space to compute by using the EM algorithm , keeping P(w|z) invariable. After that , the conditional probability is P(w|q) and P(w|s) are computed in the retrieved document according to the formulae :

$$P(w|q) = \sum_{z \in Z} P(w|z)P(z|q)$$

$$P(w|s) = \sum_{z \in Z} P(w|z)P(z|s)$$

For detailed understanding refer : A Lecture on PLSA

4 Proposed Approach

We plan to use first of all parse the user query in order to extract useful words and extract the useful words from them. Then a Wikipedia search is done using the extracted words from the user query. Then from the numerous documents generated we will take first , say N documents (N not decided yet) and apply PLSA on them . After applying PLSA answer ranking algorithm is applied on the output generated by PLSA . This will finally give the best n (n according to the user requirement) sentences that are relevant to the question asked by the user as the output. We plan to carry out the coding in JAVA and at present plan to use the existing **Natural Language processing Toolkit**

References

[1] A Web Knowledge Based Approach for Complex Question Answering
Han Ren , Donghong Ji , Chong Teng , Jing Wan
2011. http://www.springerlink.com/content/9646253882343438/

[2] Modeling Semantic Question Context for Question Answering : Proceedings of the Twenty-Second International FLAIRS Conference (2009)
Protima Banerjee and Hyoil Han
2009

[3] Probabilistic Latent Semantic Indexing
Published in : SIGIR'99 Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval
Thomas Hoffmann http://dl.acm.org/citation.cfm?id=312624.312649

[4] Topic -Based Document Segmentation with Probabilistic Latent Semantic Analysis
Published in : CIKM'02 Proceedings of the eleventh international conference on Information and knowledge management
Thorsten Brants , Francine Chen , Ioannis Tsochantaridis
2002 http://dl.acm.org/citation.cfm?id=584792.584829

[5] Web usage mining based on probabilistic latent semantic analysis
Published in KDD '04 Proceedings of the tenth ACM SIGKDD international conferece on Knowledge discovery and data mining
Xin Jin , Yanzan Zhou , Bamshad Mobasher
2004 http://dl.acm.org/citation.cfm?id=1014076