International Journal of Science Engineering and Advance Technology

An Approach to Knowledge Discovery by Data Harvesting

[1]M.Kiran kumar,[2]B.Srinivas

1M.Tech student,2AsstProf,Dept.of CSE Srinivasa engineering college, amalapuram

merigakirankumar510@gmail.com, sriv.vasv@gmail.com

Abstract:

In mining technology the text mining plays a vital role in today's life. Text mining is cluster data like user needs and classify the data.But its having some challenges like Information is in unstructured textual form, Not readily accessible to be used by computers, Dealing with huge collections of documents.One can express personal experiences and opinions on almost anything, at review sites, forums, discussion groups, blogs ..., (called the user generated content.)They contain valuable information in this processing cost is indeed. However In text and opinion mining problems we not solved So this paper address the problem of knowledge discovery for question and answers. Here we are presented knowledge discovery with Markov techniques and comparable techniques, these are present rigorous information about the mining. My results shows potent and emotive information for asking questions.

Key words:Question Answering, knowledge discover, opinion mining, markov method.

Introduction:

Each and every annum internet users are increased approx. 14-16 percent[1],[2] and we get billion dollars on internet. In these type of internet we need not to provide absolute information so our previous work tells after some years these income generation is reverse down. So we concentrate on knowledge mining. Two main types of information on the Web[4]. Facts and OpinionsCurrent search engines search for facts (assume they are true).Facts can be expressed with topic keywords [3]. Search engines do not search for opinions.Opinions are hard to express with a few keywords. How do people think of Motorola Cell phones? Current search ranking strategy is not appropriate for opinion retrieval/search.Word-of-mouth on the Web.One can express personal experiences and opinions on almost anything, at review sites, forums, discussion groups, blogs ..., (called the user generated content.).They contain valuable information. Web/global scaleNo longer limited to your circle of friends. Our interest: to mine opinions expressed in the user-generated content An intellectually very challenging problem.Practically very useful.Through "Community Based Question Answering" forums, people can seek answers to questions that belong to differentcategories and can also share their knowledge on any specific problem which is of interest to some other user. CommunityBased Question Answering forums give better answers to questions because unlike automated answering systems, they arebased on human intelligence.A huge amount of question and answer pairs have been accumulated in the repositories over the years. For example –WikiAnswers – one of the most well known hosts more than 13 million well answered questions in 7000 different categories (asof 2011). In the World Wide Web era, a comparison activitytypically involves: search for relevant webpages containing information about the targetedproducts, find competing products, read reviews,and identify pros and cons. In this paper, we focuson finding a set of comparable entities givena user‟s input entity.

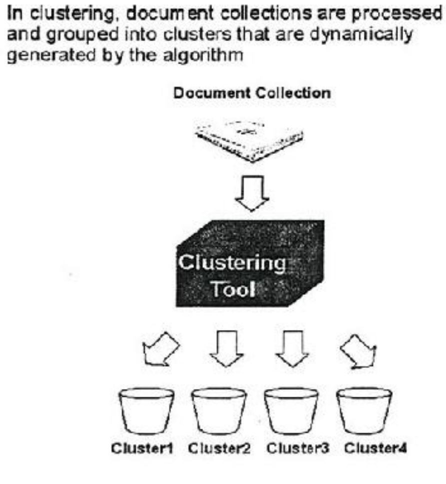
Related Work:

The existing cQA forums mostly support only textual answers. Unfortunately, textual answers may not provide sufficient natural and easy-to-grasp information. The answers are described by long sentences which generally makes it very tedious to interpret. Clearly, it will be much better if there are some accompanying videos and images that visually demonstrate the process or the concept. In the existing system users usually post URLs that link to supplementary images or videos in theirtextual answers. Therefore we can conclude that in a way the existing cQA forums do not provide adequate support in usingmedia information. Our effort on comparator mining is associated to theinvestigate on entity and relative extraction in information extraction [9]. Jindal and Liu [10], [11] also proposed a comparator mining methods for mining relative sentences and relationships. Both class and sequential rules learned to annotate the result of news and review domain to mine relative sentences as well as relationship. The similar methods followed by author [10] also applied to comparative question identification. Though, their methods characteristically can accomplish elevated precision but endure from low recall [11]. Bootstrapping methods have been shown to be very effective in previous information extraction research (Riloff, 1996; Riloff and Jones, 1999; Ravichandran and Hovy, 2002; Mooney and Bunescu, 2005; Kozareva et al., 2008).

Proposed Work:

In this work Normalized forms of dates, numbers, …Allows applications to use information very easily.Abstracts from different morphological variants of a single term likeThe canonical name is the most explicit, least ambiguous name constructed from the different variants found in the document.Reduces ambiguity of variants.So we are using clustering and classification methods,Partitions a given collection into groups of documents similar in contents, i.e., in their *feature vectors*.Two clustering enginesHierarchical Clustering tool,Binary Relational Clustering tool.Both tools help to identify the topic of a group by listing terms or words that are common in the documents in the group.Thus, provides overview of the contents of a collection of documents.

Our proposed application will give answers for the questions in any one of the following media ormats as selected by the user based on the question he/she enters: (a) Only text: It means that the original textual answers are sufficient (b) Text + image: It means that image information needs to be added (c) Text + video: It means that only video information needs to be added (d)Text + image + video: It means that we add both image and video information As per the design we have proposed an algorithmic approach for selecting the accurate video, image and text for the corresponding answers We have named it as “Multimedia answer generation from web information”.


In clustering, document collections are processed and grouped into clusters that are dynamically generated by the algorithm
Document Collection
Clustering Tool
Cluster1
Cluster2
Cluster3
Cluster4

Assign documents to preexisting categories (“topics” or “themes”).Categories are chosen to match the intended use of the collection.categoriesdefined by providing a set of sample documents for each category.

Advantages

Proposed method considers the diverse ranking is also important to enriched media data.It finds the relevant Diverse Search of Social Images for multimedia data.

Conclusion and Future Work

Existing system uses a novel scheme to answer questions using media data by leveraging textual answers in cQA. For a given QA pair, our scheme first predicts which type of medium is appropriate for enriching the original textual answer.

Following that, it automatically generates a query based on the QA knowledge and then performs multimedia search with the query. Proposed diverse relevance ranking scheme for social image search, which is able to simultaneously take relevance and diversity into account. It leverages both visual information of images and the semantic information of tags. Finally, query-adaptive reranking and duplicate removal are performed to obtain a set of images and videos for presentation along with the original textual answer.

In our future work, will further improve the scheme, such as developing better query generation method and investigating the relevant segments from a video.

References:

[1] S. A. Quarteroni and S. Manandhar, “Designing an interactive open domain question answering system,” J. Natural Lang. Eng., vol. 15, no. 1, pp. 73–95, 2008.

[2] D. Mollá and J. L. Vicedo, “Question answering in restricted domains: An overview,” Computat. Linguist., vol. 13, no. 1, pp. 41–61, 2007.

[3] H. Cui, M.-Y.Kan, and T.-S. Chua, “Soft pattern matching models for definitional question answering,” ACM Trans. Inf. Syst., vol. 25, no. 2, pp. 30–30, 2007.

[4] R. C. Wang, N. Schlaefer, W. W. Cohen, and E. Nyberg, “Automatic set expansion for list question answering,” in Proc. Int. Conf. Empirical Methods in Natural Language Processing, 2008.

[5] L. A. Adamic, J. Zhang, E. Bakshy, andM. S. Ackerman, “Knowledge sharing and Yahoo answers: Everyone knows something,” in Proc. Int. World Wide Web Conf., 2008.

[6] G. Zoltan, K. Georgia, P. Jan, and G.-M.Hector, Questioning Yahoo! Answers, Stanford InfoLab, 2007, Tech. Rep.

[7] H. Yang, T.-S.Chua, S. Wang, and C.-K.Koh, “Structured use of external knowledge for event-based open domain question answering,” in Proc.ACM Int. SIGIR Conf., 2003.

[8] T. Yeh, J. J. Lee, and T. Darrell, “Photo-based question answering,” in Proc. ACM Int. Conf. Multimedia, 2008.

[9] G. Li, R. Hong, Y.-T. Zheng, S. Yan, and T.-S. Chua, “Learning cooking techniques from youtube,” in Proc. Int. Conf. Advances in Multimedia Modeling, 2010.

[10] L. Nie, M. Wang, Z. Zha, G. Li, and T.-S. Chua, “Multimedia answering: Enriching text QA with media information,” in Proc. ACM Int. SIGIR Conf., 2011.