An Automated Approach to Comparative Question Identification and Comparator Extraction

S MD SARTAJ BASHA, NARESH ACHARI

Student of M.Tech Shri Shirdi Sai Institute Of Science And Engineering Vadiyampeta, Anantapuramu

Assistant Professor, Shri Shirdi Sai Institute Of Science And Engineering Vadiyampeta, Anantapuramu

Abstract Comparing entities are an important part of decision making process. To assist decision making it is useful to compare entities that share common utility but have distinguishing peripheral features. One possible approach is comparable entity mining from comparative questions. The technique used is weakly supervised bootstrapping approach which identifies the comparative question and extract the comparable entity. This is done by detecting whether a given question is comparative or not. A sequential pattern is generated and is called an indicative extraction pattern (IEP) if it can be used to identify comparative questions and extract comparator pairs with high reliability. This method achieves the F1-measure of 82.5 percent in comparative question identification and 83.3 percent in comparable entity extraction. In proposed system Clique grow analysis is used in which the comparable relations are extended and it is used to manipulate various query logs from users. Ranking method is used to rank the comparable entities for user' input entity and the results show highly relevance to user's comparison intent.

Keywords: *Bootstrapping, Comparable Entity mining, information extraction, IEP, sequential pattern mining.*

.

I. INTRODUCTION

Comparisons are one of the convincing ways of evaluation. For example in the business environment whenever a new product comes into market, the product manufacturer wants to know consumer opinions on the product and how the product compares with those of its competitors. Extracting such information can significantly help businesses in their marketing and product benchmarking efforts. Clearly the product comparisons are not only useful for product manufacturers but also to potential customers as it enable customers to make better purchasing decision.

This type of comparison activity is very common in daily life but requires high knowledge skill. To make better decisions it is probably an attempt to compare entities that the purchasers are exciting in. In this paper comparative questions and comparators are defined as follows. Comparative question is a question that intends to compare two or more entities. It has to mention these comparable entities explicitly in the question. Comparator is an entity which is a target of comparison in a comparative question. The bootstrapping framework needs ranking and selection procedures to guide learning process According to these definitions, Q1 is not comparative question while Q2 is comparative question and nokia and Samsung are comparators.

Q1. "Which one is better?"
Q2. "What's the difference between nokia and Samsung?"

The goal of this work is mining comparators from comparative questions and furthermore provides and rank comparable entities for user's input entity

appropriately. The comparative question has to be a question with intent to compare at least two entities. The question containing at least two entities is not a comparative question if it does not have an comparison intent. A weakly supervised bootstrapping method is used to identify comparative questions and extract comparators simultaneously. The weakly supervised method achieves 82.5 percent F1-measure in comparative question identification, 83.3 percent in comparator extraction and 76.8 percent in end-to-end comparative question identification and comparator extraction. The comparator mining results can be used for a commerce search or product recommendation system.

A. Bootstrapping method

The bootstrapping method is also called self-training, is a form of learning that is designed to use even less training examples, therefore sometimes called weakly-supervised.

Bootstrapping starts with a few training examples, trains a classifier, and uses thought-to-be positive examples as yielded by this classifier for retraining. As the set of examples grows, the classifier improves, provided that not too many negative examples are misclassified as positive, which could lead to deterioration of performance. A great advantage of bootstrap is its simplicity. Weakly supervised method is a pattern-based approach and aims to learn the sequential patterns which can be used to identify comparative question and extract comparators simultaneously. It is a straightforward way to derive estimates of standard errors and confidence intervals for complex estimators of complex parameters of the distribution. Moreover, it is a suitable way to control and check the durableness of the results. Bootstrapping is a method for assigning the accuracy.

B. Information extraction

Information extraction is the task of automatically extracting structured information from unstructured readable documents. For purchasing a product a wealth of information that can be very helpful in accessing the comparable entities and opinions toward products. Almost every day people are faced with a situation that must decide upon one thing or the other. To make better decisions probably attempt to compare entities that the customer are interesting in. These days many web search engines are helping people look for their interesting entities. Therefore a comparison mining system, which can automatically provide a summary of comparisons between two entities from a large quantity of web documents, would be very useful in many areas such as marketing.

The work is divided into two tasks to effectively build a comparison mining system. First classify the sentences into comparatives and non-comparatives and the second is related to comparative mining.

C. Ranking method

The possible comparators are ranked for a user's input. A comparator would be more interesting for an entity if it is compared with the entity more frequently. A simple ranking function Rfreq(c,e) which ranks comparators according to the number of times that a comparator c is compared to the user's input e in comparative question archive as

$$Rfreq(c,e) = N(Qc,e), \quad (1)$$

Where Qc,q is a set of questions from which c and e can be extracted as a comparator pair. This ranking function is called as Frequency-based Method. Another ranking method called Reliability-based Method is the pattern that is selected to extract comparator pair from question in comparator mining phase. Though frequency is efficient for comparator ranking, the frequency-based method can suffer when an input occurs rarely in question collection. In this case, the Frequency-based method may fail to produce a meaningful ranking result. Then representability should be considered. If a

comparator is compared to many other important comparators which can also be compared to the input entity, it would be considered as a valuable comparator in ranking.

II. RELATED WORK

"Mining Comparative Sentences and Relations", Comparison is a way to help users explore alternatives, i.e., helping them to make a decision among comparable items. Comparator mining is related to the research on entity and relation extraction in information extraction. The most relevant work is mining comparative sentences and relations. Their methods applied class sequential rules and label sequential rules to identify comparative sentences and extract comparative relations. Accuracy is less and it has low recall value. Bootstrapping method has been shown to be very effective in information extraction research. "Learning Surface Text Patterns for a Question Answering System", Surface text patterns can be learned it explore the power of surface text patterns for open domain question answering system. It uses the pattern learning algorithm to learn patterns and measure accuracy. "The Page Rank Citation Ranking: Bringing Order to the Web", The web is huge and extremely diverse in terms of content, quality and structure. The most relevant pages of the user's query be ranked at the top. The link structure of the web is producing ranking of every web page known as PageRank. PageRank could be used to separate out a small set of commonly used documents which can answer most queries. The full database only needs to be consulted when the small database is not adequate to answer a query. "Identifying Comparative Sentences in Text Documents", The comparative sentences is identified in text documents. The problem is related to but quite different from sentiment/opinion sentence identification or classification. Comparison can both be subjective or objective. It first categorizes comparative sentences into different types and then presents a novel integrated pattern discovery and supervised learning approach to identifying comparative sentences from text documents. In existing system bootstrapping method is used to learn sequential pattern. A sequential pattern is defined as a sequence where it can be a word, a POS tag, or a symbol denoting either a comparator or the beginning or the end of a question.

III. PROPOSED MODEL

In Proposed system Clique Grow analysis is used, it is a graph based approach used to manipulate various query logs from users. It aims to improve extraction pattern application and mine rare extraction pattern. It is used to identify the ambiguous entities. It provide results based on user logs and profession information. This method is used to ensure high precision and high recall value and is used to predict transitivity of known comparable relations. Automatic suggestion of comparable entities can assist users in their comparison activities before making their purchase decisions.

A. Modules

Comparative question analysis: A question is a linguistic expression used to make a request for information, or the request made using such an expression. Questions assessing comparative judgments are often phrased as directed comparisons, that is, product1 is to be compared to product2. The comparative questions have keywords such as difference, vs and so on. Admin analyze the questions whether it is comparative or not. IEP implementation: This defines the sequential patterns and identifies the starting and ending of the sentences. A sequential pattern is called an indicative extraction pattern if it can be used to identify comparative questions and extract comparators with high reliability. Then formally define the reliability score of a pattern. Once a

question matches an IEP, it is classified as a comparative question and the token sequences corresponding to the comparator slots in the IEP are extracted as comparators. Pattern evaluation: The bootstrapping algorithm is pattern based approach used to analyze the seed points. It is used to extract the features based on the seed points. Evaluate the patterns with several features. New comparator pairs are extracted from the question collection using the latest IEPs. The new comparators are added to a reliable comparator repository and used as new seeds for pattern learning in the next iteration. The process iterates until no more new patterns can be found from the question collection. Decision making: Given a question, select the longest one among patterns which can be applied to the question. Provide the decision making based on pattern evaluation. If a comparator is compared to many other important comparators which can be also compared to the input entity, it would be considered as a valuable comparator in ranking.

Performance Evaluation: It evaluate the good comparative question identification pattern and extract the good comparators and a good comparator pair should occur in good comparative questions to bootstrap the extraction and identification process. Then calculate FP rate to check whether the decision making is correct or not.

Advantages of Proposed System

- Improve recall and precision values.
- Dissolve the ambiguity in entities by using comparator analysis with patterns.
- Expand the aliases name.

IV. ARCHITECTURE

In figure i) first the input query is given by the user. It analyze whether the given question is comparative question

or not. If the given question is comparative question then the comparator pairs are extracted. Bootstrapping approach is used to identify the comparative question and extract comparator simultaneously. The features of the comparator pairs are extracted. IEP is indicative extraction pattern, once a question matches an IEP it is classified as comparative question and the token sequence corresponding to the comparator slots in the IEP are extracted as comparators. From the comparative question and the comparator pair all possible sequential patterns are generated and evaluated by measuring the reliability score.

Clique grow analysis is used, it defines the multiple patterns. If a comparator is compared to many other important comparators which can also be compared to the input entity, it would be considered as a valuable comparator in ranking. The best comparator are detected and decisions are made on analyzing the features of an entity.

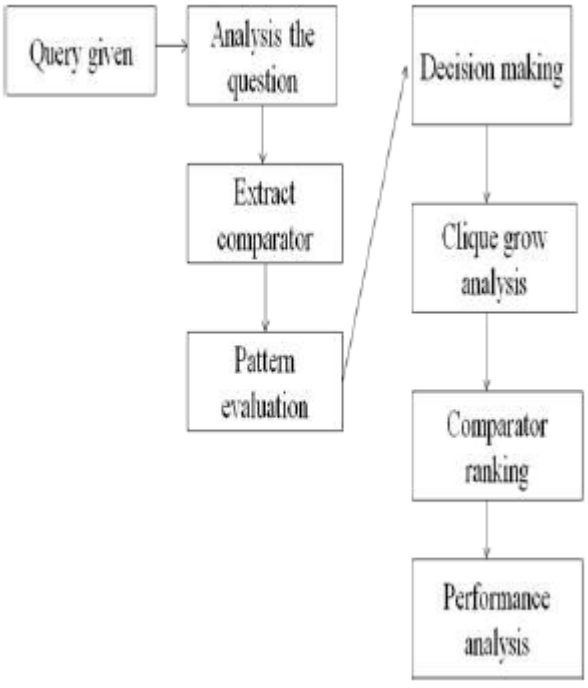

Figure i) System Architecture

VI. CONCLUSION

This paper proposed clique grow analysis that allows multiple patterns to be included and compared to its opponent for better purchase decision making. The comparable entities are extracted from the comparative question which is obtained from the user query and the features of an entities are extracted. The best comparator is detected and comparator ranking method is implemented for analyzing all the products and to find their competitors. It is used to dissolve the ambiguity in entities. The results of comparator mining can be used for commerce search or product recommendation system.

REFERENCES

[1 [1] Califf .M.E and Mooney .R.J, "Relational Learning of Pattern-Match Rules for Information Extraction," Proc. 16th Nat'lConf. Artificial Intelligence and the 11th Innovative Applications of Artificial Intelligence.

[2] Cardie .C, "Empirical Methods in Information Extraction,"Artificial Intelligence Magazine, vol. 18, pp. 65-79.

[3] Gusfield .D, Algorithms on Strings, Trees, and Sequences:Computer Science and Computational Biology. Cambridge Univ. Press,1997.

[4] Haveliwala.T.H, "Topic-Sensitive Pagerank," Proc. 11th Int'l Conf. 2002.

[5] Jeh.G and Widom.J, "Scaling Personalized Web Search," Proc. 12th Int'l Conf. World Wide Web (WWW '02), pp. 271-279, 2003.

[6] Jindal.N and Liu.B, "Identifying Comparative Sentences in Text Documents," Proc. 29th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '06), pp. 244-251, 2006.

[7] Jindal.N and Liu.B , "Mining Comparative Sentences and Relations," Proc. 21st Nat'l Conf. Artificial Intelligence (AAAI '06), 2006.

[8] Kozareva.Z, Riloff.E , and Hovy.E , "Semantic Class Learning from the Web with Hyponym Pattern Linkage Graphs," Proc. Ann. Meeting of the Assoc. for Computational Linguistics: Human Language Technologies (ACL-08: HLT), pp. 1048-1056, 2008.

[9] Li.S, Lin.C.-Y, Song.Y.-I, and Li.Z , "Comparable Entity Mining from Comparative Questions," Proc. 48th Ann. Meeting of the Assoc. for Computational Linguistics (ACL '10), 2010.

[10]Linden.G, Smith.B, and York.J, "Amazon.com Recommendations: Item-to-Item Collaborative Filtering," IEEE Internet Computing, vol. 7, no. 1, pp. 76-80, Jan./Feb. 2003.

[11] Mooney.R.J and Bunescu.R , "Mining Knowledge from Text Using Information Extraction," ACM SIGKDD Exploration Newsletter, vol. 7, no. 1, pp. 3-10, 2005.

[12] Page.L , Brin.S , Motwani.R , and Winograd.T, "The PagRank Citation Ranking: Bringing Order to the Web," Stanford Digital Libraries Working Paper

[13] Radev.D, Fan.W, Qi.H ,Wu.H, and Grewal.A , "Probabilistic Question Answering on the Web," J. Am. Soc. for Information Science and Technology, pp. 408-419, 2002.

[14] Ravichandran.D and Hovy.E , "Learning Surface Text Patterns for a Question Answering System," Proc. 40th Ann. Meeting on Assoc. for Computational Linguistics (ACL '02), pp. 41-47, 2002.

[15] Riloff.E and Jones.R , "Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping," Proc. 16th Nat'l Conf. Artificial Intelligence and the 11th Innovative Applications of Artificial Intelligence Conf. , pp. 474-479.