Getting Better Alternatives by Comparable Entity Mining: A Survey

Priyanka R. Shirsath, Ujwala M. Patil

P. G. Student, Dept. of Computer Engineering, RCPIT, North Maharashtra University, MS, India

Associate Professor, Dept. of Computer Engineering, RCPIT, North Maharashtra University, MS, India

ABSTRACT: As internet is a user friendly technology by using it people solve their so many of problems such as shopping, study and a lot more. Also, people uses internet whenever they wants to buy some products, use some services or wants to go out for vacation on some places. But searching for relevant option become time consuming process as there are several links available even for the single thing. So it becomes more difficult for people to decide one best option among so many of them as search engines works on general terms. For this reason, this paper focuses specifically on showing the best option. For this, a weakly supervised bootstrapping approach is used which aims to identify comparative questions and evaluate comparable entities concurrently. For the persons who does not know the another choices for anything they want buy or use this method provides the best solution. The weakly supervised bootstrapping approach is very useful for knowing the better alternatives in people, products, places, any applications or any services.

KEYWORDS: Comparable Entity Mining, Information Extraction, Bootstrapping, Sequential Pattern Mining.

I. INTRODUCTION

Comparison becomes most useful when we want to know the advantages and disadvantages of products such as laptop, computer or services such as medical treatments. For example, when someone is interested in particular place like Shimla, Kerala for vacations or in products like mobile phones, cars, bikes for purchasing, he/she would want to know the alternatives before reaching the final decision. In this world people easily compare two or more things but only when they know everything about those entities. For example, choosing between car or bike for going out for college is easy as we know a little bit about these transport. That means comparison involves a high knowledge. Consumer Reports, PC Magazines and Online Media such as CNet.com are the sources that provide such a high knowledge of comparison and try to satisfy the need of comparison.

 In the past few years, a significant amount of research was done on sentiment and opinion extraction and classification. While this paper focuses only on identifying comparative questions and mine comparators. For example, if someone is interested in buying laptop like Lenovo and if he posts a question about it then Dell, Apple and HP will be shown as a result for the laptop depending on the ranking of comparators which is used in [1].

The comparison becomes difficult when entities having different functionalities which may create another problem. For example, one might compare "iPhone" with "Nokia Lumia" as a mobile phone or compare "iPhone" with "PSP" as a portable game device [1]. The comparison becomes complicated when an entity have different identities. For Example, people do compare "Paris versus London" as a location and "Paris versus Nicole" as a celebrity.
As comparison is very crucial to make decision, plenty of questions are posted online. For Example, "Which is the best car BMW or Farari?" this is the comparative question. There can be a question having two or more entities but may not have comparison intention. For example, "How do I change a .bmp image to a jpeg?"Therefore, the comparative question and comparators are defined as follows:

- **Comparative Question.** It is a question having comparison intention with two or more entities. But this question has to mention these entities explicitly.

- **Comparator.** It is a posted entity which is a target of comparison in comparative question.

According to this definition a comparative question has to be a question with a comparison intention, comparing at least two objects. For example,

Q1. "Which one is better?"
Q2. "Is Renault Lodgy the next rising car?"
Q3. "What is the difference between LED Television and LCD Television?"

From above definition,Q3. is the comparative question and the "LED Television", "LCD Television" are the comparators. So, from above definition the questions Q1. and Q2. are not comparative question.

The goal of this work is to detect the comparative question and excerpt comparators simultaneously. For this a weakly supervised bootstrapping approach is developed. The resultant comparators are shown as result by ranking them. The result will be more useful for the people who do not know the alternative options for the entities in which they are interested.

II. RELATED WORK

Table 1. Related Work

Sr. no.	Author	Title	Year	Approach	Remark
1	Shasha Li et. al	Comparable Entity Mining from Comparative Questions	2013	Presented an approach which can identify comparative questions specifically and extract comparators concurrently.	Use WH-questions only means cannot identify the question as comparative question which are starting by verbs.
2	Z. Kozareva, et al.	Semantic Class Learning from Web with Hyponym Pattern Linkage Graphs	2008.	Presented a novel approach to weakly supervised semantic class learning from the web, using a single powerful hyponym pattern joint with graph-structures	There were errors created by their algorithm which were initiated by incorrect proper name extraction
3	N. Jindal and B. Liu	Mining Comparative Sentence and Relations	2006	Their methods applied CSRs and LSRs learned from marked corpora to identify comparative sentences and extract comparative relations	Methods typically can attain high precision but agonize from low recall.
4	Kennedy, C.	Comparatives, semantics of. In Enclycopedia of Language and Linguistics,	2005.	Discusses the gradability of comparatives and measure of gradability.	The semantic analysis is centered on logic, which is not directly applicable to finding comparative question and extract comparators.
5	Dave et al.	Mining the Peanut Gallery:	2003.	Experiments a number of learning methods for review	The classifiers performs well on

|					
		Opinion extraction and semantic classification of product reviews.		classification.	whole reviews, but unwell on sentences because a sentence contains much less information.
6	Pang, B. et al.	Thumbs up? Sentiment Classification Using Machine Learning Techniques.	2002.	Examines several supervised machine learning methods	Methods were inspected for sentiment classification of movie reviews
7	D. Ravichandran and E. Hovy,	Learning Surface Text Patterns for a Question Answering System	2002.	Developed a method for learning surface text patterns automatically.	Only those questions were used which do not have a long answers because it was affecting precision of patterns
8	E. Rilof and R. Jones	Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping	1999.	Presented a multilevel bootstrapping algorithm that engenders both the semantic lexicon and extraction patterns instantaneously.	The algorithm produces high-quality dictionaries for several semantic categories
9	Moltmann, F.	Coordination and comparatives.	1987.	Studies semantic and syntax of comparative sentences, but uses only limited vocabulary.	It is not able to do task of ascertaining comparative question and comparators extraction concurrently.

III. METHODOLOGY
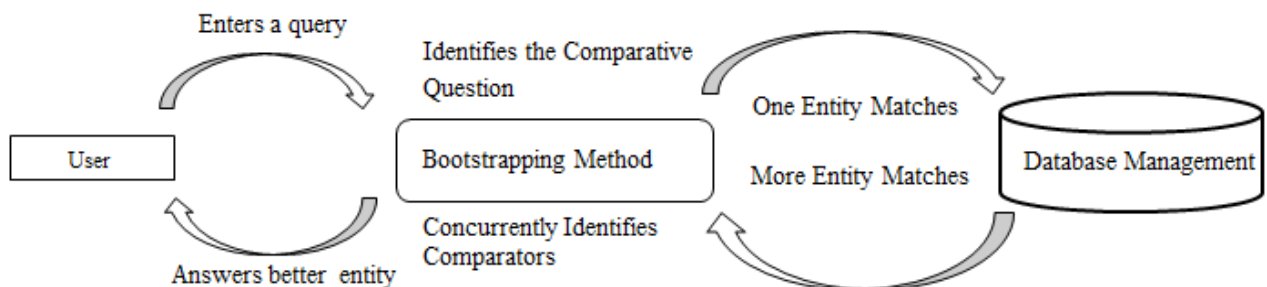
Fig.1.System Overview

The weakly supervised bootstrapping method is a pattern-based approach which aims to learn sequential pattern further which can be used to identify comparative question and extract comparators at the same time rather than using separate Class Sequential Rules[CSR] and Label Sequential Rules[LSR] [3,4].

A. *Description of the Weakly Supervise Bootstrapping Method:*

Aim of the proposed method is to find better option by identifying comparative question and mining comparable items at the same time. This method consists of three main steps.

Step 1: Pattern Generation:

The algorithm generates the sequential pattern by using three kinds:

- Lexical pattern: Replaces comparators with $C symbols and has #start and #end symbols at the beginning and ending respectively.
- Generalize pattern: Replaces comparators with POS tags [4] and have #start and #end symbols at the beginning and ending respectively.
- Specialize pattern: It replaces all the words by POS tags [4] in comparator slotand have #start and #end symbols at the beginning and ending respectively.

Step 2: Pattern Evaluation:

To generate IEPs, the sequential patterns have to be reliable one whose reliability score can be calculated as follows:It requires three factors,

I.   Reliability Score$R1 = \dfrac{\text{No. of sequences containing initial seed comparator pair.}}{\text{All the questions containing same comparator pair.}}$

II.   Support S= No. of sequence patterns containing same comparator pair.

III.   Lookahead Reliability Score $R2 = \dfrac{\text{No. of sequences containing initial seed comparator pair whose S} > 3.}{\text{All the questions containing same comparator pair.}}$
R2 is zero if S < 3.

Therefore, calculating Final reliability score $R_{final}=\lambda (R1) + (1- \lambda) R2$, where $\lambda=0.5$

The sequential patterns with high reliability score are regarded as IEPs and used to extract new comparator pair as stated in algorithm.

Step 3: Ranking of Comparators.

After extracting the comparators ranking is done. For this it uses two methods:

- Comparability-Based Ranking Method: It ranks high that entity which is more frequently compared with other entity [1].
- Graph-Based Ranking Method: When an entity is not frequently compared with other entity then representability can be considered. Here to rank the comparator by representability PageRank is used [1].

IV. PSEUDO CODE

Step 1: Generate the initial seed pattern

Step 2: Extract initial comparator pairs.

Step 3: Check the question collection for those questions containing initial comparator pair and regard them as comparative question.

Step 4: From comparative question and comparator pair. All the potential sequential patterns are generated and evaluated by measuring their reliability score.

Step 5: Newly evaluated sequential patterns are IEPs and from them new comparator pairs are extracted from question collection.

Step 6: The new comparators are added to the reliable comparator repository and used as new seeds for pattern learning in the next iteration.

Step 7: Repeat step 3 to 5 until no more new sequential patterns can be found from question collection.
Step 8: go to step 1.
Step 9:End

V. CONCLUSION

The Weakly Supervised Bootstrapping Method focuses precisely on identifying comparative question and comparable entities. So, it is very useful for the users who do not know the alternatives for the things they are interested in, which they have posted online. This method mines the comparable items from comparative questions and gives better entity as well as alternative option for the online posted entity. It can effectively be used to detect comparative question and mine comparator pair concurrently instead of using separate Sequential Rules. It significantly improves recall in both tasks that is identifying comparative question and extracting comparators while maintains high precision. It can be used for commerce search or product recommendation system.

REFERENCES

1. S. Li, C.-Y. Lin. Y-I. Song, and Z. Li., "Comparable Entity Mining from Comparative Questions", IEEE Transactions on Knowledge and Data Engineering, Vol. 27, Issue No. 7, 2013.
2. Z. Kozareva, E. Rilof, and E. Hovy, "Semantic Class Learning from the Web with Hyponym Pattern Linkage Graphs", Proc. Ann. Meeting of the Assoc. forComputational Linguistics: Human LanguageTechnologies (ACL-08: HLT), pp. 1048-1056, 2008.
3. Nitin J. and B. Liu, "Mining Comparative Sentences and Relations", Proc. 21st Nat'l Conf. Artificial Intelligence (AAAI '06), 2006.
4. Nitin J. and B. Liu, "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.
5. Kennedy, C., "Comparatives, semantics of InEnclycopedia of Language and Linguistics", Second Edition, Elsevier, 2005.
6. Dave, K. Lawrence, S., and Pennock, D., "Mining the Peanut Gallery: Opinion extraction and semantic classification of product reviews", WWW'*03*, 2003.
7. Pang, B., Lee, L., and Vaithyanathan, S., "Thumbs up? Sentiment Classification Using Machine Learning Techniques", EMNLP'02, 2002.
8. D.Ravichandran and E. Hovy., "Learning Surface Text Patterns for a Question Answering System", Proc. 40th Ann. Meeting on Assoc. for Computational Linguistics (ACL '02), pp. 41-47, 2002.
9. E. Rilof and R. Jones, "Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping", Proc. 16th Natl Conf. Artificial Intelligence and the 11th Innovative Applications of Artificial Intelligence Conf. (AAAI '99/IAAI '99), pp. 474-479, 1999.
10. Moltmann F.,"Coordination and comparatives", Ph.D. Dissertation, MIT. Cambridge Ma, 1987.