International Journal of Emerging Trend in Engineering and Basic Sciences (IJEEBS)
ISSN (Online) 2349-6967 *Volume 2 , Issue 1(Jan-Feb 2015), PP1-5*

Research on Comparable Entity Mining from Comparative Questions

Kanchan Kundgir[1], Poonam Dere[2], Surabhi Mohite[3], Priti Mithari[4]
[1 2 3 4] *(Computer & Dr. D.Y. Patil Institute of Engineering and Technology, India)*

Abstract :- Making Comparisons between things is a typical part of human decision making process. But however, it is difficult to know what are to be compared and what can be the alternatives. For eg., if someone is interested in certain products such as digital cameras, then he /she would want to know what the alternatives are and compare different cameras before making any purchase. This type of comparison activity is very common in our daily life but requires high knowledge skill inorder to make much better choice. Therefore, to address this difficulty, we are presenting a novel way to automatically mine comparable entities from comparative questions that users posted online. In this paper, we focus on finding a set of comparable entities provided a user's input entity. For example, provided an entity like Nokia N95 (a mobile phone), we want to find comparable entities such as Nokia N82, Blackberry and so on. To ensure high precision and high recall, we are developing a weakly-supervised bootstrapping method for comparative question identification and comparable entity extraction by leveraging a large online question archive. The results will prove to be very useful in helping users' exploration of alternative choices by suggesting comparable entities based on other users' prior requests.

Keywords: - Bootstrapping method, Comparable entity mining, Information extraction, Part Of Speech Tags, sequential pattern mining.

I. INTRODUCTION

In decision-making process, comparing alternative options is one of the necessary steps that we carry out daily. But this activity requires high knowledge expertise to make better choice. For instance, while doing shopping of a laptop one must have detailed knowledge of its specifications like Processor, Storage, Graphics, Memory, Display, etc. In such cases, it becomes difficult for an individual with insufficient knowledge to make a good decision on which laptop to buy and also comparing the alternative options for the same.

Magazines such as PC Magazine, Consumer Reports and online media like CNet.com which make efforts in providing editorial comparison content and surveys. The comparison activity in the World Wide Web normally involves- search for applicable web pages enclosing information regarding the targeted products, discovering competing products, and recognizing their pros and cons. Our focus, in this paper, is on finding a set of comparable entities provided a user's input entity. For e.g., provided an entity like Nokia N95 (mobile phone), we would want to find entities that are comparable like iPhone, Blackberry, Nokia N82, HTC and etc.. Inorder to extract comparable entities from relative matter, we first need to find out whether the question is relative or not.

1.1. Terms and concepts

1.1.1. Information Extraction: The process of automatically drawing out structured information from unstructured and/or a semi- structured machine-readable document is called Information Extraction.

Methods for information extraction:

1. Rule-based

2. Pattern based

3. Supervised learning

1.1.2. Sequential Pattern mining**:** Sequential Pattern mining is mainly concerned with finding statistically relevant patterns between data examples where the values are delivered in a sequence.

1.1.3. Comparable entity mining**:** Comparable entity mining is concerned with extracting the comparable entities from the text/questions/web corpus.

1.1.4. POS Tags (Part-of-speech)**:** Part-of-speech of a word is a linguistic category defined by its syntactic or morphological behaviour. Common POS categories are: noun, verb, adverb, adjective, pronoun, interjection, preposition and conjunction. Then there are many categories which arise from different forms of these categories. POS tags used in this paper are:

NN: Noun, NP: noun phrases, NNP: Proper Noun, NNS: Noun plural, PRP: Pronoun, VBZ: Verb, Present tense, Third person singular, JJR: Comparative Adjective, JJS: Superlative Adjective, CC: coordinating conjunction (and, or) and WP: wh-pronoun (who, what).

Our effort on comparable entity mining is related to the study on entity and relation removal in information extraction. According to our definition, a comparative question has to be a query with intention to contrast at least two entities. We exploit this insight and develop a weakly supervised bootstrapping means to identify comparative questions and extract comparable entities at the same time.

1.2. Comparative questions: A question whose purpose is to compare two or more entities and these entities are explicitly mentioned in the question.

1.3. Comparator: An entity in a comparative question which is to be compared [1].

According to the definitions, Q1 & Q2 below are not comparative questions whereas Q3 is. "Mumbai" and "Pune" are comparators.

Q1. "Which one is better?"
Q2. "Is Pune the best city?"
Q3. "Which city is better Mumbai or Pune?"

The results will be very useful in helping users' exploration of alternative choices by suggesting them comparable entities based on other previous users' requests.

II. RELATED WORK

2.1. Overview

In terms of discovering related items for an entity, their work is similar to the research on recommender systems, to recommend items to a user. Recommender systems is similar between items and/or their statistical correlations in user log data [2]. For example, Amazon recommends products to its customers based on their own previous purchase; similar customers' previous purchase, and similarity between products. a comparable item is not equivalent to Recommending an item for finding customer item. In Amazon, the purpose of recommendation is to entice their customers to add more items to their shopping carts by suggesting similar or related items.

In the case of comparison, they help users explore alternatives, i.e., helping them make a decision among comparable items. For example, it is reasonable to recommend "iPod speaker" or "iPod batteries" if a user is interested in "iPod," but they are not compare them with "iPod." However, items that are comparable with "iPod" such as "iPhone" or "PSP" which were found in comparative questions posted by users are difficult to be predicted simply based on item similarity between them. Although they are all music players, "iPhone" is mainly a mobile phone, and "PSP" is mainly a portable game device. They are similar but also different therefore beg comparison with each other. It is clear that comparator mining and item recommendation are related but not the same. Their comparator mining is related to then research on entity and relation extraction in information extraction [2], [4], [5].

2.2 . Supervised Comparative Mining Method

Major contribution of Jindal and Liu on mining comparative sentences and relations, in their system used class sequential rules (CSR) and label sequential rules (LSR). CSR maps a sequence pattern S(s1s2.....sn) to class C. Class C is either comparative or non-comparative question .and LSR maps an input sequence pattern S(s1s2...si...sn) to a labeled sequence S'(s1s2...si...sn)by replacing one token si in the input sequence with a designated label (li). This token is referred as the anchor.

J&L work on this method and treated comparative sentence identification as a classification problem and comparative relation extraction is called as an information extraction problem. They first manually created a set of 83 keywords is similar to the indicators of comparative sentences. These keywords were then used as pivots to create part-of-speech (POS) sequence data.

The Table 1 below shows brief view of the Literature Survey.

TABLE I
LITERATURE SURVEY

Sr. no	Paper Name	Conference	Approaches	Advantages	Disadvantages
1	Identifying Comparative Sentences in Text Documents.	ACM SIGIR Conf. Research and Development in Information Retrieval, 2006.	Combination of class sequential rule (CSR) mining and machine learning [6].	Extract comparative sentences from text is useful for many applications	It can achieve high precision but suffer from low recall.
2	Mining Comparative Sentences and Relations	Artificial Intelligence (AAAI '06), 2006.	Identify comparative sentences from the texts and to extract comparative relations to its identified comparative sentences [7].	Evaluating an entity or event is to directly compare it with a similar entity or event.	It can achieve high precision but gives low recall.
3	Comparable Entity Mining from Comparative Questions	Proc. 48th Ann. Meeting of the Assoc. for Computational Linguistics (ACL '10), 2010.	Mining the comparators from given entities of comparative questions [3].	Identifies comparative questions and extracts that comparators simultaneously using one single pattern	Their rules achieved high precision but low recall.
4	Relational Learning of Pattern Match Rules for Information Extraction	Proc. 16th Nat'l Conf. Artificial Intelligence and the 11th Innovative Applications of Artificial Intelligence (AAAI '99/IAAI '99), 1999.	Desired information can be extracted from natural language texts [4].	It can be research on relation and entity extraction in information extraction	The learned patterns employ limited syntactic and semantic information to identify potential slot fillers and their surrounding context.
5	Mining Knowledge from Text Using Information Extraction.	ACM SIGKDD Exploration Newsletter, vol. 7, no. 1, pp. 3-10, 2005.	Information extraction extracts structured data or knowledge from unstructured text [8].	Information Extraction is extracting structured data from unstructured or semi-structured web pages.	Cannot reduce demanding corpus-building requirements of information system.
6	Learning Surface Text Patterns for a Question Answering System	Proc. 40th Ann. Meeting on Assoc. for Computational Linguistics (ACL '02), pp. 41-47, 2002	Automatically learning such regular expressions from the web, for given types of questions [9].	Their system assumes each sentence to be a simple sequence of words and searches for repeated word orderings as evidence for useful answer phrases.	The system does not classify or make any distinction between upper and lower case letters.

III. PROPOSED SYSTEM

3.1 Weakly Supervised Method For Comparator Mining

Jindal and Liu proposed method for Comparator mining. In this method CSR and LSR rules are used respectively [5]. CSR is a classification rule. It maps a sequence pattern S ($s_1s_2.....s_n$) to a class C. In our problem, C is either comparative or non-comparative. And LSR is a labeling rule. It maps an input sequence pattern S($s_1s_2...s_i...s_n$) to a labeled sequence S'($s_1s_2...l_i...s_n$) by replacing one token si in the input sequence with a designated label (l_i). However, their methods typically can achieve high precision but suffer from low recall. So, to overcome this problem, we develop a weakly supervised bootstrapping method , Our weakly supervised method is a pattern-based approach similar to J&Ls method, but it is different in many aspects. Instead of using separate CSRs and LSRs, our method aims to learn sequential patterns which can be used to identify comparative question and extract comparators simultaneously.

In our approach, a sequential pattern is defined as a sequence S ($s_1s_2...s_i...s_n$) where si can be a word, a POS tag, or a symbol denoting either a comparator ($C), or the beginning (#start) or the end of a question (#end). A sequential pattern is called an indicative extraction pattern (IEP) if it can be used to identify comparative questions and extract comparators in them with high reliability.

3.2 Mining Indicative Extraction Patterns

Our weakly supervised method is based on two assumptions:

1) If a sequential pattern can be used to extract many reliable comparator pairs then it is very likely to be an IEP.
2) The pair is capable to compare if a comparator pair can be extracted by an IEP.

Based on these two assumptions, system design boot- strapping algorithm. The bootstrapping process starts with a single IEP. We extract a set of initial seed comparator pairs from it. For each comparator pair all questions containing the pair are retrieved from a question collection and regarded as comparative questions. From comparative questions and comparator pairs all possible sequential patterns are generated and evaluated by measuring their reliability score defined later in the Pattern Evaluation section. Patterns evaluated as reliable are IEPs and are added into an IEP repository.

There are two key steps in this method:
1) Pattern generation
2) Pattern evaluation

The three kinds of sequential patterns are generated from sequences of questions are as follows [3]:
i) Lexical patterns- Lexical patterns indicate sequential patterns consisting of only words and symbols ($C, #start, and #end) .
ii) Generalized patterns- A lexical pattern can be too specific. So we generalize lexical patterns by replacing one or more words their POS tags.
iii) Specialized patterns- we perform pattern specialization by adding POS tags to all comparator slots. For example, from the lexical pattern '<$C or $C>' and the question 'ipod or zune?', '<$C=NN or $C=NN?>' will be produced as a specialized pattern.

We evaluate all candidate patterns and select patterns whose score is more than threshold as IEPs. All necessary parameter values are empirically determined.

3.3 Comparator Extraction

By applying learned IEPs, we can easily identify comparative questions and collect comparator pairs from comparative questions existing in the question collection. Given a question and an IEP, the details of the process for comparator extraction are shown as follows:

1. ‘‘‘‘Generate sequence for the comparative question .If the IEP is a pattern without generalization, we just need to tokenize the questions and the sequence is the list of resulted tokens. Otherwise, phrase chuncking is necessary. The sequence is a list of resulted chunks. Take the question “Which is better phone iPhone or nokia n95 ?” for example. If we apply the pattern “Which is better $C or $C?” to the question, we generate a sequence “which|is|better| phone|iphone|or|nokia|n95|?”. If we apply pattern “Which is better NN $C or $C?” to the question, a sequence “Which/WP | is/VBS| better/ JJR|phone/NN|iphone/NP|or/CC|nokia n95/ NP|?/?”.
2. Check whether sequence of the question matches the given pattern. If IEP is a specialized pattern, the POS tag sequence of extracted comparators should follow the constraints specified by the pattern.
According to above observation, system examined the following strategies:

3.1.1 Random strategy

Given a question, randomly select a pattern among patterns which can be applied to the question.

3.1.2 Maximum length strategy

Given a question, select the longest one among patterns which can be applied to the question. According to the discussion above, the longer the pattern is, the more tokens in the question can be exactly covered which means that the pattern is more suitable for the question.

3.1.3 Maximum reliability strategy

Given a question, select the most reliable one among patterns which can be applied to the question.

IV. COMPARATOR RANKING

The remaining issue is to rank possible comparators for a user's input. The following ranking models are examined for this issue.Comparability-Based Ranking Method comparator is more interesting for an entity if it is compared with the entity more frequently. system define a simple ranking function R 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 Q.

V. GRAPH BASED RANKING METHOD

frequency is efficient for comparator ranking, the frequency-based method can suffer when an input occurs rarely in question collection; for example, suppose the case that all possible comparators to the input are compared only once in that questions. the Frequency-based method is fail to produce a meaningful ranking result In this case. Then, Representability should also be considered. System regard a comparator representatiive if it is frequently used as a baseline in the area the user is interested in. For example, when one wants to buy a smart phone and he/she is considering "Nokia N82," "Nokia N95" is the first one he/she wants to compare. That's because "Nokia N95" is a well-known smart phone and it's usually used as a baseline to help users know the performance of other smart phones better.

One possible solution to consider represent ability can be to use graph-based method such as PageRank. 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. Based on this idea, system examine PageRank algorithm to rank comparators for a given input entity which combine frequency and represent ability.

ALGORITHM

Algorithm 1 Weakly-Supervised Model

Input: CP, G
Initialize solution: $Q \leftarrow \{\}, P \leftarrow \{\}, P_{new} \leftarrow \{\}, CP_{new} \leftarrow CP$
1. **Repeat**
2. $P \leftarrow P + P_{new}$
3. $Q_{new} \leftarrow ComparativeQuestionIdentify(\ CP_{new})$
4. $Q \leftarrow Q\ +\ Q_{new}$
5. **for** $q_i \in G$ **do**
6. **if** $IsMatchExistingPatterns(P, q_i)$ **then**
7. $Q \leftarrow Q - q_i$
8. **end if**
9. **end for**
10. $P_{new} \leftarrow MineGoodPatterns(Q)$
11. $CP_{new} \leftarrow \{\}$
12. **for** $q_i \in G$ **do**
13. $cp \leftarrow ExtractComparableComparators(P, qi)$
14. **if** $cp \neq NULL$ **and** $cp \notin CP$ **then**
15. $CP_{new} \leftarrow\ CP_{new} + \{cp\}$
16. **end if**
17. **end for**
18. **until** $P_{new} = \{\}$
19. **return** P

Fig. 1. Pseudocode of the bootstrapping algorithm.

VI. CONCLUSION

In this paper, we present a new supervised method for identifying comparative questions and extraction of comparator pairs at the same time. We rely on the key insight that a good comparative question identification pattern should extract good comparator pairs, and a good comparator pair should occur in good comparative questions to bootstrap the extraction and identification process. This method considerably improves recall in together tasks whilst maintain elevated precision.

Comparator mining outcome can be useful for commerce exploration or product recommendation organization. For instance, automatic proposition of comparable entities can help out users in their assessment activities earlier than building their acquire decision. In addition, the outcome can make available helpful information to companies which would like to recognize their competitors.

ACKNOWLEDGEMENTS

The authors wish to thank the researchers, publishers for making their resources available and teachers who provided them valuable assistance to the writing of the research summarized. Finally, they extend their heartfelt gratitude to their friends for assisting in the collection of the topics and their family for the help and inspiration extended.

REFERENCES

[1] K. Kundgir, P. Dere, S. Mohite, P. Mithari, "A Study on Comparable Entity Mining from Comparative Questions", International Journal of Advancement in Engineering Technology, Management and Applied Science, vol. 1, no. 2349-3224, 2014, 62-66.
[2] 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. 25, no. 7,* 2013, 1498-1509.
[3] S. Li, C.-Y. Lin, Y.-I. Song, and Z. Li, "Comparable Entity Mining from Comparative Questions," *Proc. 48th Ann. Meeting of the Assoc. for Computational Linguistics* (ACL '10), 2010.
[4] M.E. Califf and R.J. Mooney, "Relational Learning of Pattern- Match Rules for Information Extraction," *Proc. 16th Nat'l Conf. Artificial Intelligence and the 11th Innovative Applications of Artificial Intelligence* (AAAI '99/IAAI '99), 1999.
[5] C. Cardie, "Empirical Methods in Information Extraction," *Artificial Intelligence Magazine, vol. 18, 1997, 65-79.*
[6] N. Jindal 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), 2006, 244-251.
[7] N. Jindal and B. Liu, "Mining Comparative Sentences and Relations," *Proc. 21st Nat'l Conf. Artificial Intelligence* (AAAI '06), 2006.