ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 12, December 2014

A Review on Comparable Entity Mining

Shrutika Narayane, Sudipta Giri

M.E. Student, Department of Information Technology, MIT, College of Engineering, Kothrud, Pune, India

Professor, Department of Information Technology, MIT, College of Engineering, Kothrud, Pune, India

ABSTRACT: Comparisons of different things plays crucial role in human decision making process. Even though, decision making is common in our daily life but requires proper knowledge and skills to know what should get compare and what will be alternatives so as to make good decision. In this paper, we review the background and state-of-the-art of comparable entity data mining based on comparative questions.We first introduce the general background of entity mining from comparative questions and review related phases, such as information extraction and comparator ranking. With each phase, we provide a related background, discuss the technical challenges, and review current research on the techniques used in that phase. This survey is concluded with a discussion of latest experimental results from research articles.

KEYWORDS: comparable entity mining; comparators; inductive extraction pattern; bootstrapping algorithm.

I. Introduction

It is human nature to try to compare things. Laptops, mobiles, iPhones, cars etc. are compared on number of features. In decision-making process, comparing alternative options is one of the necessary steps that we carry out in day-to-day life. However it requires skillful and high knowledge expertise person. In today's era everyone is using World Wide Web (WWW) and it is obvious to compare things online. For e.g. for online laptop shopping user must have detailed knowledge of its specifications such as processor, memory, storage, graphics, display, etc. In such case, it becomes difficult for a person with insufficient knowledge to make a good decision to finalize best laptop according to his/her need and also making comparison of alternative options available in market. Comparative question and its comparators are two main components of decision making process.

Comparative questions: A question with purpose of comparing two or more entities which are explicitly mentioned in the question archived by online users.

Comparator: Target entities in a comparative question which are to be compared are comparative entities or called as comparators.

In the following example Q1 & Q2 are not comparative questions whereas Q3 is comparative question in which "BMW" and "Skoda" are comparators.

Q1. "Which one is better?"
Q2. "Is BMW the best car?"
Q3. "Which car is better carBMW or Skoda?"

The outcomes of these comparative questions will be very useful in helping user's exploration i.e. recommending variousalternatives choices by suggesting comparable entities on the basis of other previous online user's requests.

The procedure of discovering related items for an entity is similar to recommender system, which recommends items for users. Recommender systems mainly rely on similarities between items and/or their statistical correlations in user log data. In literature we can found many research articles focusing on comparator mining [1], [2], [3],[4]. In our paper, we tried to make inside in to their proposed techniques with their pros and cons.

The rest of paper is organized as follows. In Section II a short literature survey is given, Section III gives a very brief review of information extraction. Latest experimental results are presented in section IV with conclusion in section V.

II. RELATED WORK

The work on comparator mining is related tothe research on entity and relation extraction ininformation extraction. Specifically, the most relevant work is by Jindal and Liu [1], [15]on mining comparative sentences and relations. Their

10.15680/ijircce.2014.0212018

methods applied class sequential rules (CSR) and label sequential rules (LSR) learned from annotated corpora to identify comparative sentences and extract comparative relations respectively in the news and review domains.Bootstrapping methods have been shown to be very effective in previous information extraction research [6], [8]. Bootstrapping technique is used to extract entities with a specific relation.

III. Information Extraction

IInformation Extraction (IE) deals with locating specific pieces of data in natural-language documents, thereby mining structured and meaningful information from unstructured and/or a semi-structured one is called as Information Extraction [5]. One type of IE, named entity recognition, involves identifying references to particular kinds of objects such as names of people, companies, and locations. There are mainly three methods used for information extraction as [6], [7], [8] given below,

1. *Rule based Extraction:*

One approach of IE is to automatically learn pattern-based extraction rules for identifying each type of entity or relation. For example, the system developed by Rapier in [9]. Patterns are expressed in an enhanced regular-expression language; and a bottom-up relational rule learner is used to induce rules from a corpus of labeled training examples. Inductive Logic Programming (ILP) [10] has also been used to learn logical rules for identifying phrases to be extracted from a document [11], [12].

2. *Pattern based extraction:*

Pattern based approaches build on annotated text fragments (the patterns), where words/phrases are labeled with linguistic information, e.g. POS-tag, word lemma, or syntactic information. Those patterns are matched against linguistically annotated text to detect relationships [13].

3. *Supervised Learning:*

Supervised learning is the machine learning task of inferring a function from labeled training data. The training data consist of a set of training examples. In supervised learning, each example is a pair consisting of an input object (typically a vector) and a desired output value (also called the supervisory signal). However, supervised training of accurate entity and relation extractors is costly, requiring a substantial number of labeled training examples for each type of entity and relation to be extracted. Because of this, many researchers have explored semi-supervised learning methods that use only a small number of labeled examples of the predicate to be extracted, along with a large volume of unlabeled text [14]. All the above information extraction methods can be used for comparator methods as in [2], [3], [4] by Li Shasha, Jindal and Liu in [1],[15].

A. *Design Considerations:*

Supervised comparative mining method was proposed by Jindal and Liu [1], [15] which is a baseline for comparison. It focuses mainly on two rules mentioned as Class Sequential Rule (CSR) & Label Sequential Rule (LSR) as descried below.

a. *Class Sequential Rule (CSR):*

It is a classification rule which maps a sequence pattern S ($s_{1,}$ s_2 . . . s_n) (a class C. C is either comparative or noncompetitive). Every CSR is associated with two parameter support and confidence.

b. *Label Sequential Rule (LSR):*

It maps an input sequence pattern S ($s_{1,}$ s_2 . . . s_i . . . s_n) to a labeled sequence S ($s_{1,}$ s_2 . . . l_i . . . s_n) by replacing token s_i in the input sequence with a designated label (l_i) and this token is referred as the anchor.

Jindal and Liu [1] method have been proved effective in their experimental setups. However, it has the some drawbacks as given below,

- The performance of Jindal and Liu’s method depends mainly on a set of comparative sentence indicative keywords [3].
- Users can express comparative sentences or questions in many different ways. To have high recall, a large annotated training corpus is necessary. This is an expensive process
- CSRs and LSRs introduced by Jindal and Liu in [15] mostly a combination of POS tags and keywords. It is a surprise that their rules achieved high precision but low recall.

10.15680/ijircce.2014.0212018

B. *A Weakly Supervised Method for Comparator Mining:*

To resolve the conflict in extracting comparative questions and its comparator with high precision as well as with high recall a Weakly Supervised Bootstrapping method is introduced by Li Shasha in [2].

1. *Indicative Extraction Patterns Mining:*

Indicative Extraction Pattern (IEP) is a sequential pattern which can be used for identification of comparative questions along with comparator extraction with high reliability. A question is classified as a comparative question if it matches an IEP and the token sequences corresponding to the comparator slots in the IEP are extracted as comparators. If a question matchesmultiple IEPs, the longest IEP is used. Therefore, instead of manually creating a list of indicative keywords, we create a set of IEPs automatically, referred as weakly supervised method which is iterative as shown in Fig. 1. The two key steps in this algorithm are pattern generation and pattern evaluation.

2. *Pattern Generation:*

The weakly supervised IEP mining is highly based on two key assumptions as [3], [16], [17]

- If a sequential pattern can be used to extract many reliable comparator pairs, it is very likely to be an IEP.
- If a comparator pair can be extracted by an IEP, the pair is reliable.

Based on these key assumptions, bootstrapping algorithm designed as shown in Fig. 1.

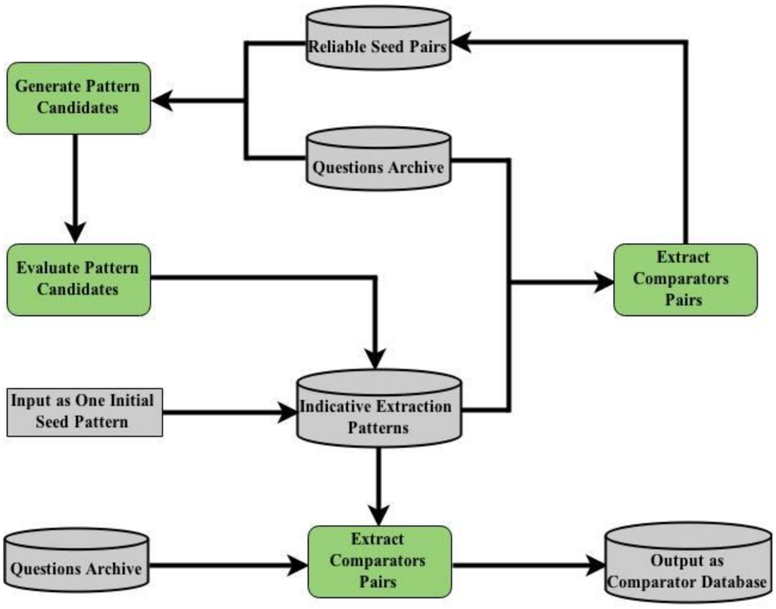
Reliable Seed Pairs
Generate Pattern Candidates
Questions Archive
Evaluate Pattern Candidates
Extract Comparators Pairs
Input as One Initial Seed Pattern
Indicative Extraction Patterns
Questions Archive
Extract Comparators Pairs
Output as Comparator Database

Fig. 1. Flow chart of the bootstrapping algorithm

To generate sequential patterns, Li Shasha in [1],[3] used surface text mining method introduced in [2]. In this method, comparators in the question are replaced by symbol $Cs in any given comparative question and its comparator pair. The symbol #start is attached to the beginning of the each sentence and the symbol #end at the end of sentence. Li Shasha in [3] used some heuristic rules and phrase chuncking for diversity reduction of sequence data and mine potential patterns. Following three kinds of sequential patterns can be generated from sequences of questions as:

a. *Lexical Patterns:*

These patterns indicate sequential patterns consisting of only words and symbols ($C, #start, and #end).

b. *Generalized Patterns:*

A lexical pattern is too specific for matching. So lexical patterns are generalized by replacing one or more words their POS tags.

c. *Specialized Patterns:*

Pattern specialization is done by adding POS tags to all comparator slots. For example, from the lexical pattern '<$C or $C>' and the question 'Paris or London?', '<$C=NN or $C=NN?>' will become specialized pattern.

10.15680/ijircce.2014.0212018

Note that in this method, lexical patterns are used to generate generalized patterns and the combined set of generalized patterns and lexical patterns are used to generate specialized patterns [1],[3].

3. *Pattern evaluation:*

Bootstrapping gives very few reliable comparator pairs in its early stage. Hence for discovering more reliable pair's, pattern evolution operation is performed. In this case, the value might be underestimated which could affect the effectiveness of on distinguishing IEPs from non-reliable patterns. This problem is mitigated by a look-ahead procedure. The next step is to rank possible comparators for a user's input [1], [3].

C. *Comparator Extraction:*

By employing IEPs, it is easy to identify comparative questions and collect comparator pairs from available data. For given question and an IEP, comparator extraction process is described in [1], [2], [3], [4] as follows:

1. *Generate sequence for the comparative question:*

If the IEP is a pattern without generalization, then tokenize the questions and the list of resulted tokens is the sequence. Otherwise, phrase chuncking is needed. The sequence is a list of resulted chunks.

2. *Check whether sequence of the question matches with the given pattern:*

If IEP is a specialized pattern, the POS tag sequence of extracted comparators should follow the constraints specified by the pattern.

However, a result of [3] shows about 67 % comparative questions can match to multiple patterns, and from 11 % comparative questions, we can extract different comparator pairs. Li Shasha in [3], [4] examined three different strategies to solve the issue of comparator extraction.

D. *Comparator Ranking:*

The comparability and graph based methods are examined rank possible comparators for user's input [1], [3], [18] which are described below,

1. *Comparability-Based Ranking Method:*

Frequent comparison of entity with particular entity would make comparator more interesting. Based on this intuition, a simple ranking function $R_{freq}(c;e)$ ranks comparators on the basis of number of times that a comparator c is get compared to the user's input e in online comparative question archive *Q*.

$$R_{freq}(c;e) = N(Q_{c,e}) . \qquad eq.\ (1)$$

Where (Q_c, e) is a set of questions from which comparator *c* and user input e can be extracted as a comparator pair. This method also known as frequency based Method. The another ranking function is R_{rel} by combining reliability scores estimated in comparator mining phase

$$R_{rel}(c;e) = \sum_{q \in Q_{c,e}} R(p_{q,c,e}) . \qquad eq.\ (2)$$

Where *p q, c, e* means the pattern that is selected to extract comparator pair of c and e from question [3].

2. *Graph Based Ranking Method:*

Frequency is consider as efficient parameter for comparator ranking but the frequency-based ranking method [3] can suffer when an user input occurs rarely in collection of questions; for example,suppose all possible comparators to the input are compared only once in questions. In this case, this method may fail to results correct ranking result. Hence in addition to it representing ability should also be considered. We regard a comparator representative if it is frequently used as a baseline while making comparison of interested entity.

Graph based page rank method is one of the solutions to get ability. A comparator can be considered as valuable comparator in ranking if it is compared to too many other important comparators including the input entity. Based on this idea, Page Rank algorithm is examined to rank comparators for a given input entity, which combine frequency and represent ability [3].

IV. EXPERIMENTAL RESULTS

In this section the experimental results of different research papers are compared and discuss.

A. *Comparative Question Identification and Comparator Extraction:*

The latest experimental results on comparative question identification and comparator extraction for the data of 60M questions mined from Yahoo! Answers' question title field can be found in [4]. The experimental results of Li

10.15680/ijircce.2014.0212018

Shasha [3] compared with Jindal and Liu's [1] methods are shown in Table I. In the Table I, column with *Identification* only shows the performances in comparative question identification, column with *Extraction* only shows the performances of comparator extraction when only comparative questions are used as input, and last column with *All* shows the end-to-end performances when question identification results were used in comparator extraction.

In terms of precision, the method described in [1] is competitive to method used in [3] in comparative question identification. However, the recall is significantly lower in [1] than [3]. In the end-to-end experiments, weakly supervised method of [3] performs significantly better than method of [1]. F1-measure of [1] in All is about 30 % and 32 % worse than the scores of Identification only and Extraction only respectively, our method only shows small amount of performance decrease (approximately 7-8 %).

TABLE I. COMPARISON BETWEEN EXPERIMENTAL RESULTS OF LI SHASHA [3] AND JINDAL AND LIU [1]

	Identification Only SET A + SET B			**Extraction Only SET B**		**ALL SET B**		
	Jindal and Liu (CSR)[1]		**Li Shasha [3]**	**Jindal and Liu (LSR)[1]**	**Li Shasha [3]**	**Jindal and Liu [1]**		**Li Shasha [3]**
	SVM	NB				SVM	NB	
Recall	0.601	0.537	**0.817***	0.621	**0.760***	0.373	0.363	**0.760***
Precision	0.847	**0.851**	0.833	0.861	**0.916***	0.729	0.703	**0.776***
F-score	0.704	0.659	**0.825***	0.772	**0.833***	0.493	0.479	**0.778***

Here * indicate statistically significant improvements over Jindal and Liu (CSR) SVM or Jindal and Liu (LSR) according to t-test at p < 0:01 level performed by Li Shasha.

B. *Ranking Results of Comparability-BasedvsGraph-Based Ranking Methods:*

The proposed algorithm is consists of three main steps. Ranking results of comparability and graph based ranking methods of [3] are shown in Table II. For some queries whose comparator's frequency differs significantly, such as "iphone" and "BMW 328i" the ranking results of two methods do not make many differences. That's because frequency plays the main role in the ranking process for these queries in graph-base ranking methods. However, for queries whose comparators share similar frequency, such as "BMW 328i","Nokia N75" and "Nikon D200" the differences between two methods are obvious. These experimental results show both graph based and page rank methods are effective for both comparative question identification and comparator extraction.

TABLE II. RANKING RESULTS OF COMPARABILITY-BASED VS GRAPH-BASED RANKING METHODS FROM LI SHASHA [4]

Rank	**iphone**		**BMW 328i**	
	Comparability	**PageRank**	**Comparability**	**PageRank**
1	ipod touch	ipod touch	Toyotoavalon	Cadillac Cts
2	Blackberry	Blackberry	BMW 338i	Toyotoavalon
3	Itouch	itouch	Audi A3	Accura TL
4	Storm	storm	Honda Accord A08	Honda accord A08
5	Voyager	voyager	Accura TL	Audi A3

V. CONCLUSION

This paper surveys various research articles and there experimental results that are currently available and discusses the pros and cons for each of them. A thorough comparison between different methods based on experimental results for 60M questions. After surveys it is found that a novel weakly supervised method described in [1], [2], [3],[4] identifies comparative questions and extracts comparator pairs simultaneously with high precision and high recall. The results of [2], [3], [4] can be used for a commerce search or product recommendation system from user comparison interest. For example, automatic suggestion of comparable entities can assist users in their comparison activities and will help for better purchase decisions.

REFERENCES

1. NitinJindal and Bing Liu,'Identifying Comparative Sentences in Text Documents',Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 244-251, 2006.
2. Li Shasha, Chin-Yew Lin, Young-In Song, and Zhoujun Li,'Comparable Entity Mining from Comparative Questions', Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL'10), 2010.
3. Li Shasha, Chin-Yew Lin, Young-In Song, and Zhoujun Li,'Comparable Entity Mining from Comparative Questions', Knowledge and Data Engineering, IEEE Transactions on 25, no.7, pp. 1498-1509, 2013.
4. LiShasha, Chin-Yew Lin, Young-In Song, and Zhoujun Li,'Comparative Entity Mining', U.S. Patent no. 8, 484, 201, July 2013.
5. Califf M. Elaine and Raymond J. Mooney,'Relational Learning of Pattern Match Rules for Information Extraction',Proceedings of the 16thNational Conference on Artificial Intelligence and the 11th Innovative Applications of Artificial Intelligence Conference (AAAI '99/IAAI '99), pp. 328-334, 1999.
6. Mooney J. Raymond and RazvanBunescu, 'Mining Knowledge from Text Using Information Extraction', ACM SIGKDD explorations newsletter 7.1, pp. 3-10, 2005.
7. Cardie Claire, 'Empirical Methods in Information Extraction',Artificial Intelligence Magazine, vol. 18, pp. 65-79, 1997.
8. Riloff Ellen and Rosie Jones, 'Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping', Proceedings of the 16th National Conference on Artificial Intelligence and the 11th Innovative Applications of Artificial Intelligence Conference (AAAI '99/IAAI '99), pp. 474-479, 1999.
9. Califf M. Elaine and Raymond J. Mooney, 'Bottom-up Relational Learning of Pattern Matching Rules for Information Extraction', Journal of Machine Learning Research, vol 4, pp. 177–210, 2003.
10. Mooney J. Raymond and Loriene Roy, 'Content-based Book Recommending using Learning for Text Categorization', Proceedings of the 5th ACM Conference on Digital Libraries, pp. 195–204, 2000.
11. FreitagDayne, 'Toward General-purpose Learning for Information Extraction', Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and COLING-98 (ACL/COLING-98), pp.404–408, 1998.
12. StephenSoderland, 'Learning Information Extraction Rules for Semi-Structured and Free Text', Machine Learning, vol. 34, nos. 1-3, pp. 233-272, 1999.
13. Chang Chia-Hui and Shao-Chen Lui, 'IEPAD: Information Extraction Based on Pattern Discovery', Proceedings of the 10th international conference on World Wide Web (WWW' 01), 2001.
14. Carlson Andrew, Justin Betteridge, Richard C. Wang, Estevam R. HruschkaJr, and Tom M. Mitchell, 'Coupled Semi-supervised Learning for Information Extraction', Proceedings of the 3rd ACM international conference on Web search and data mining, pp. 101-110, 2010.
15. Nitin Jindal and Bing Liu, 'Mining Comparative Sentences and Relations', Proceedings of the 21st National Conference on Artificial Intelligence (AAAI '06), vol. 22, pp. 1331-1336. 2006.
16. RiloffEllen, 'Automatically Generating Extraction Patterns from Untagged Text', Proceedings of the 13th National Conference on Artificial Intelligence, pp. 1044-1049, 1996.
17. RadevDragomir, Weiguo Fan, Hong Qi, Harris Wu, and AmardeepGrewal, 'Probabilistic Question Answering on the Web', Journal of the American Society for Information Science and Technology, 56, no. 6, pp. 571-583, 2005.
18. Haveliwala H.Taher, 'Topic-Sensitive Pagerank', Proceedings of the 11th International Conference on World Wide Web (WWW' 02), pp. 517-526, 2002.

10.15680/ijircce.2014.0212018