IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 8, August 2016

A Survey on Comparable Entity Mining from Comparative Questions by Using Weakly Supervised and Markov-Logic Network

P. Ragha Vardhani[1], **Y. Indira Priyadarshini**[2]

Asst. Professor, Dept. of CSE, Ravindra College of Engineering for Women, Kurnool, Andhra Pradesh, India[1, 2]

Abstract: Comparing two objects is a very typical part for human decision making process. However, this process is not always easy to know what to compare and what are the substitutes. To address this difficulty, we propose a novel way to automatically mine comparable entities from comparative questions that users posted online. To get high precision and high recall, we implemented a weakly-supervised bootstrapping method, to achieve comparable entity extraction and comparative question identification by leveraging a large online question archive. The experimental results show our proposed method achieves F1- measure of 82.5% in comparative question identification and 83.3% in comparable entity extraction. The comparative question identification and comparable entity extraction significantly outperform an existing state-of-the-art method.

Keywords: Automatically mine comparable entities, weakly-supervised bootstrapping method, F1- measure of 82.5%

I. INTRODUCTION

A comparison of alternative options is a major step in deciding to carry out every day. For example, someone interested in some products such as digital cameras, and know what it is that alternatives before making a purchase I'd like to compare different cameras. This type of activity is very common in everyday life, but require high knowledge and skills. Magazines such as PC Magazine and Consumer Reports and online media such as CNet.com attempt in providing editorial comparison content and surveys to satisfy this need. In the web the users are search the product, tagged product, finding the compare products, read reviews and identify the pros and cons. In this paper, we study on finding a set of comparable entities given a user's input entity. For example: Let an entity be Nokia N95 (a mobile phone), we want to find comparable entities such as Nokia N82, iPhone, iPod and so on.

In general, for a variety of reasons that people do apples and oranges comparison, since it is difficult to decide whether the two sides are not comparable. For example, “BMW” and “Ford” might be comparable as “automobile manufacturers” or as “market segments that their products are targeting”, but we hardly ever see people comparing “BMW 328i” (car model) and “Ford Focus”. Things get more complicated when an entity has several uses. For example, one might compare “PSP” and “iPhone” as “portable game player” while compare “Nokia N95” and “iPhone” as “mobile phone”. Fortunately, sufficient comparative questions are posted online, which provide confirmations for what people want to compare, example: “What to buy iPhone or iPod?”.
Here we identify “iPhone” and “iPod” in this example as comparators. In this paper, we define comparators and comparative questions as:

- **Comparator:** It is an entity which is a target Comparative compared.
- **Comparative question:** Intend to compare two or more assets, and these assets has mentioned explicitly mentioned question.

According to this definition is Q1 and Q2 are not comparative questions below. Q3 is "Zune HD" and "iPod Touch" comparators.
Q1: “Which one is god?”
Q2: “Lumix GH-1 Is the best camera?”
Q3: “What is the difference between iPod Touch and Zune HD?”
The aim of this study comparators mining comparative questions. Results are comparable to other users on the basis of prior request by suggesting the existence of alternative options will be very useful in helping user's discoveries.

Comparator mines comparative questions, we must first detect if a question is comparative or not. According to our definition, a comparative question has to be a question with the intention of comparing at least two entities. Note that a query that contains at least two entities is not a comparative question if you have no intention of comparison. However, it is observed that a question is likely to be a comparative question if it contains at least two entities. We follow this idea and develop a bootstrap method to identify weakly supervised comparators compare issues and draw simultaneously. To our knowledge, this is the first attempt to specifically address the problem of finding good comparators to support the comparison of user activity. We are also the first to propose the use of comparative questions posted online that reflect what users really care about as the medium

DOI 10.17148/IJARCCE.2016.5871

from which we extract comparable entities. Our method weakly monitored reaches 82.5% F1-identification measure comparative question, 83.3% in comparison extraction, and 76.8% at the end to end identification question comparative extraction comparator exceeding method the state of the relevant art by Jindal and Liu (2006b) significantly.

The next section discusses the previous works. Section 3 presents our weakly-supervised method for mining comparison. Section 4 reports the evaluations of our techniques, and we conclude the paper and discuss future work in Section 5.

II. RELATED WORK

In terms of discovering the elements related to an entity, our work is similar to research on recommendation systems, which recommend items to a user. Recommender systems are mainly based on the similarities between the elements and / or statistical correlations in the user registration data (Linden et al.). For example, Amazon recommends products to customers based on their own purchase histories, customers buy similar stories, and the similarity between the products. However, the recommendation of an item that is not equivalent to find a similar item. For Amazon, the objective of the recommendation is to attract customers to add more items to your shopping cart by suggesting similar or related items. While in the case of the comparison, we would like to help users to explore alternatives, i.e. help you make a decision within comparable objects.

For illustration, it is reasonable to recommend “iPod batteries” or “iPod speaker” if a person is interested in “iPod”, but we would not compare them with “iPod”. However, this kind of "PSP" or the comparative questions were submitted by users “iPhone”, “iPod” as compared with products based on the similarity between the items just are difficult to predict. Although all music players, "iPhone" mainly on a mobile phone, and "PSP" is mainly a portable gaming device. Therefore, I beg comparison with each other is similar but different. This comparison mining and related substances but that the proposal is clear.

Our work on the comparator mining, information extraction and relation extraction research on the items are related (Cardia, Califf and Mooney, Soderland, Radev et al., Carreras et al.). In particular, the most appropriate study on comparative sentences and relations mining is by Jindal and Liu. Their methods, class sequential rules (CSR) and label sequential rules (LSR) applied to identify comparative sentences and relations news and comment fields to extract the comparative learned from annotated corpora. Comparative question identification and comparison questions, the same techniques can be applied mining. However, their methods are generally achieved high sensitivity but low recall (Jindal and Liu) (J & L) can suffer. However, enabling users enjoy high recall queries can give you is important in our intended application scenarios. To solve this problem, effectively taking advantage of unlabeled questions a weak-supervised learning method to improve bootstrapping model.

Bootstrapping method is very effective in previous studies was shown to retrieve information (Riloff, Riloff and Jones, Ravichandran and Hovy, Mooney and Bunescu, Kozarev et al.). Our working relationship with the presence of a specific bootstrapping using the technique to extract similar to themselves in terms of methodology. However, our task of extracting assets (comparator extraction) requires not only, but also people often dismiss IE is not required comparative questions (comparative question identification) is being issued to provide different from theirs.

III. WEAKLY SUPERVISED AND MARKOV-LOGIC NETWORK COMPARABLE ENTITY MINING

Markov logic network (MLN) to representation of interweaved constraints. MLN is most important type of entity linking method with genetic material state relating. The proposed MLN is the combination of first order logic (FOL) and Markov networks with combination of NIL-filtering and entity disambiguation stages. The representation captures the background information of the familiar entities for entity disambiguation as well as consideration of entity linking in the Knowledge Base (KB) .For instance, an individual declare preserves imply be linked to a KB entry when the state has not been well-known as an NIL. The formula on KB bases are demonstrated with four keywords: constants, variables, functions, and predicates. While the constants are referred to as objects in the database entries, that related variables are denoted as x and y for particular objects. Relationship amongst the data objects are represented as predicates. A world is an obligation of reality values to everyone probable view atoms is also referred to as predicates. Knowledge Base (KB) is an incomplete requirement of a world; every particle in it is perfect(true) , false or unidentified.

A Markov Logic Network (MLN) characterizes the joint distribution of a set of variables $X = (X_1, X_2, \ldots . X_n) \in x$ as a result of factors:

$$P(X = x) = \frac{1}{z} \pi_z f_k(x_k)$$

Where every factor f_k is a non-negative purpose of a separation of the variables x_k , and Z is normalization constant.

As extended as intended for every one $P(X = x) > 0$, for everyone x the distribution can be consistently represent as a log-linear representation:

$P(X = x) = \frac{1}{z} \exp(\Sigma_i w_i\, g_i(x))$,

Where $g_i(x)$ is the features are subjective functions of the variables situation.

An MLN L is a set of pairs (F_i, w_i) , where F_i is a principle in FOL and w_i is a real numeral represent a weight. Mutually with a predetermined position of constants, it describe a Markov network, $M_{L,C}$ where contains single

DOI 10.17148/IJARCCE.2016.5871

node for every probable preparation of every predicate is shown in L. The evaluation of the node is 1 if the ground predicate is true, and 0 or else. The probability distribution in excess of probable worlds is known by

$$P(X = x) = \frac{1}{z} \exp(\Sigma_i \Sigma_j w_i g_j(x))$$

where Z is the separation function, F is the set of every one first order formula in the MLN, is the set of groundings of the i^{th} first-order formula, and $g_j(x) = 1$ if the j^{th} ground formula is true and $g_j(x) = 0$ or else.

Describe four predicates to confine the accepted questions environment information, together with question location, Question Interaction (QI), Tissue Type and Question ontology. The formula describing the relation of and hasquestionInfo and islinkedto is defined as follows:
hasquestionInfo(i, id,+sd) ⟹ islinkedTo(i, id).

At this time, can perceive that in attendance is an added parameter (+sd) indicate in hasquestionInfo.sd consequent to id locates. The "+ " details in the beyond method indicates that necessity study a split weight for every grounded variable (sd). For example, : hasquestionInfo(i, id, 0) and hasquestionInfo(i, id, 1) are specified two dissimilar weights in our MLN model following preparation.

Correlation information from knowledge base (KB) approach interacts with entity one to entity two to solve a disambiguating an entity problem. The QI information is accumulated in the backend database with correlation measure. Based on this result and candidate KB entry distribution result , the id to associated with the majority unambiguous entries is the mainly probable id to be linked to i. Additional describe the subsequent formula to confine the dependence that an entity be supposed to be linked to id_2 if one more entity have be linked to id_1 structure a correlation with id_2. Filtering the subsequent mention type persons belong to classes with the intention of are not in the database curation objective; called NILs. In linking question with gene are stored to KB Database and NIL filter apply the QI interaction to solve the entity disambiguation problem. The subsequent formula to make sure to, every time the entity is linked to a KB entry id , it be supposed to be an entity appropriate for linking,

islinkedTo(i, id) ⇒ issuitableForlinking(i)
∃w.hasWord(w) Λ QIKeyword(w)
ΛislinkedTo(i, id_1)
Λhascandidate(j, id_2)
ΛisQIPair(id_1, id_2) ⟹ islinkedTo(j, id_2) formula(1)

The steps involved in this Markov Logic Network(MLN) are defined as follows:

Input: A Markov network represents the joint distribution of a set of variables
$X = (X_1, X_2, \ldots . X_n) \in x$, L is set of pairs (F_i,w_i)

Output: Find disambiguation result (Fi,wi).
Step 1: Identify or establish the set of disambiguation pairs from using Markov Logic Network (MLN).
Step 2: Find the set of disambiguation result (F_i,w_i) where F_i a formula in FOL is and w_i is a real number represented a weight.
∃w.hasWord(w) Λ QIKeyword(w)
ΛislinkedTo(i, id_1)
Λhascandidate(j, id_2)
ΛisQICPartner(id_1, id_2) ⟹ islinkedTo(j, id_2)
formula(1)

Step 3: If it is if $(F_i,w_i) > \mathbb{C}$ then defines a Markov network, $M_{L,C}$ where contains one node for each possible grounding of each predicate appearing in L.
Step 4: The value of the node is 1 if the ground predicate is true, otherwise its value is 0
Step 5: Discover the probability distribution over possible worlds is given by,

$$P(X = x) = \frac{1}{z} \exp(\Sigma_i \Sigma_j w_i g_j(x))$$

Step 6: In the step $g_j(x) = 1$ if the jth ground is true and $g_j(x) = 0$ otherwise.
Step 7: Return the best probability result for each pairs (F_i,w_i)
Step 8: Then now apply bootstrapping procedure collection of sequence patterns is specified as S an indicative extraction pattern (IEP) ,condition it be able to be used to identify comparative questions and extract comparators in them through elevated consistency. Primary will properly describe the consistency attain of a sample. The sequence patterns is specified as S as a sequence S where si can be a word or a representation of symbol denote moreover a comparator ($\$c$), or the beginning ($\#start$) or the end of a question($\#end$).

Input: CP, G

```
$InitializesolutionQ \leftarrow \{\}, P \leftarrow \{\}\ Pnew \leftarrow \{\}CPnew \leftarrow \{\}$
$Repeat$
    $P \leftarrow P + Pnew$
    $Qnew \leftarrow compartiveQuestionidentify\ (CP\ new)$
    $Q \leftarrow Q + Qnew$
    $Forqi \in Gdo$
        $Ifismatchexistingpatterns(p, qi\ )\ then$
            $Q \leftarrow Q - qi$
        $Endif$
    $Endfor$
    $pnew \leftarrow mineGoodpatterns\ (Q)$
    $cpnew \leftarrow \{\ \}$
    $Forqi \in Gdo$
        $cp \leftarrow extractcomparablerpatterns(p, qi)$
        $Ifcp \neq NULLandcp \notin CPthen$
            $CPnew \leftarrow CPnew + \{CP\}$
        $Endif$
    $Endfor$
$UntilPnew = \{\}$
$ReturnP$
```

A. Patterns Generation and Evaluation

To produce sequential patterns, become accustomed the exterior text pattern mining technique introduced. For some specified comparative question and its pairs, questions of each comparator are replaced with representation \$Cs. Together symbols, #start and #end, are emotionally involved to the start and the end of every sentence in the question. To decrease variety of series information and extract possible patterns, expression chunking is practical. After that, the next three kinds of sequential patterns are generated beginning series of questions:

- **Lexical patterns** point toward sequential patterns containing only the representation of symbols and of only words. They generate sequential patterns using suffix tree algorithm among consideration of two constraints that is β not more than one \$C, and its occurrence in compilation be supposed to exist additional than an empirically resolute number β.
- **Generalized patterns** are able to be as well precise simplify lexical patterns by replacing one or additional words/phrases by means of their POS tags. 2n - 1 generalized patterns can be fashioned beginning a lexical pattern containing N words exclusive of \$Cs.
- **Specialized patterns** a pattern be able to universal even though a question is relative, For this cause, carry out pattern specialization by addition POS tags to all comparator slots .

According to our primary supposition, a reliability score $R^k(p_i)$ for a contestant pattern pi at iteration k might be definite as follows

$$R^k(p_i) = \frac{\sum_{\forall cp_j \in cp^{k-1}} N_Q(p_1 \rightarrow cp_j)}{N_Q(p_1 \rightarrow cp_j)}$$

Where candidate pattern p_i can extract identified consistent comparator pairs cp_j, cp^{k-1} indicates the reliable comparator pair depository accumulated awaiting the (k − 1)th iteration.

$N_Q(x)$ means the numeral of questions rewarding a condition x. The condition $p_i \rightarrow cp_j$ specifies that cp_j can be extracted from a question by applying pattern p_i whereas the condition $p_i \rightarrow *$ specifies some question containing pattern p_i .

B. Comparator Extraction

Comparator extraction used a random based strategy to perform comparator, it randomly choose a pattern amongst patterns which be able to be useful to the question. Another type of strategy is Maximum length strategy. These strategies select a maximum pattern for given a question which is able to be applied to the question comparator extraction. From the discussion above comparator extraction in this work uses a maximum length method is able to exist exactly enclosed which means that the model is additional appropriate intended for the query.

C. Comparable Ranking Methods

The major importance of comparable based ranking methods is to compare the extra attractive entity for an entity if it is compared with the entity further regularly. Based on this insight, describe a straightforward ranking function $R_{freq}(c, e)$ which ranks the comparator results corresponding to the amount of time when the comparatorc is compare toward the user's key e in relative questions collection Q:

$R_{freq}(c, e) = N(Q_{c,e})$ where $Q_{c,e}$ is a set of questions from the comparatorc is compare toward the user's key e can be extracted as a comparator couple .Describe one more ranking function R_{rel} by combination of dependability scores predictable in comparator mining stage

$$R_{rel}(c, e) = \sum_{q \in Q c,e}^{\cdot} R(pq, c, e) \square R(p_{q,c,e})$$

where $p_{q,c,e}$ way the model that is preferred to mine comparator pair of comparator c is compare toward the user's key e from question q in comparator mining phase. This ranking function determination is present denoted as Reliability-based system.

D. Graph-Based Ranking

Although regularity is well-organized for comparator ranking, the frequency-based technique can experience whilst an effort occur infrequently in question collection; for instance, understand the case that all probable comparators to the effort are compared simply on one occasion in questions. In this case, the Frequency-based method might be unsuccessful to create a significant ranking end result. Then, Representability is supposed to moreover be considered. For instance, when individual requirements to buy a smart phone and allowing for "iphone-89","iphone 87" is the primary lone he/she needs to evaluate. It uses a graph-based Page Ranking method to compare questions. If a comparator is compared to numerous additional significant comparators which are able to be moreover compared to the input entity, it would be considered as a precious comparator in ranking. Based on this scheme, examine Page Rank algorithm to rank comparators for a known input entity which merge regularity and represent ability.

IV. EXPERIMENT EVALUATION

A. Experiment Setup

1. Source Data

All experiments were conducted in nearly 60 million drawn questions from Yahoo! Answers Question Title field. The reason that we used only one title field is to clearly express the main intent of a questioner with a form of simple questions in general.

2. Evaluation Data

Two independent data sets were created for evaluation. First, 5,200 were collected sample questions

200 questions in Yahoo! Answers category3. Two annotators were asked to label each question manually as a comparative, non-comparative, or unknown. Among them,

139 (2.67%) were classified as questions comparative, 4.934 (94.88%) as comparison, and 127 (2.44%) as unknown questions that are difficult to evaluate. We call this set SET-A. Because there are only 139 questions in comparative SET-A, we create another set containing the most comparative questions. We manually constructed a set of keywords consisting of 53 words "or" and "prefer", which are good indicators of comparative questions. In SET-A, 97.4% in comparative question contains one or more keywords in the keyword set. They randomly selected 100 questions of each other Yahoo! Answers category with an additional condition that all questions have to contain at least one keyword. These questions were labeled in the same way as a SET-A, except that its comparator were also recorded. This second group of questions concerns as SET-B. It contains 853 questions and 1,747 not compare comparative questions. For comparative question identification experiments, all questions marked in SET-A and SET-B were used. Extraction experiments for comparison, we used only SET-B. All other unmarked questions (called as SET-R) were used for the formation of our weakly supervised method.As a reference method, we implemented carefully J&L's method. Specifically, CSR for comparative question identification were taken from the marked questions, then a statistical classifier was built using the rules on CSR as features. We examine both SVM and Naïve Bayes (NB) models as reported in their experiments. For extraction of the comparator, were learned LSRs SET-B and applied to the extraction of comparison. To start the process of bootstrapping, the IEP "<# start nn/ $c vs/cc nn/$ c /. # End>" was applied to all questions in the SET-R and gathered 12,194 pairs of comparison because the seeds initial. To our weakly supervised method, there are four parameters, namely, α, β, γ, and λ, need to be determined empirically. First, all possible mined candidate patterns suffix tree using the initial seed. From these patterns of candidates, applied to SET-R and we have a new set of 59410 pairs of comparison candidates. These new pairs of candidate's comparison, 100 randomly selected pair's comparison and manually classify them into reliable or unreliable comparators. Then find α that maximize accuracy without hurting recall by investigating the frequencies of the pairs in the set labeling. By this method, α is set to 3 in our experiments.

Similarly, the β and γ parameters for model evaluation threshold is set to 10 and 0.8 respectively. For the interpolation parameter λ in equation (3), simply set the value of 0.5 by assuming that two reliability scores are equally important.

As evaluation measures for comparative identification question and the extraction of comparison was used precision, recall, and F1-measure. All results were obtained from 5-fold cross-validation. Note that the method of J & L's needs training data, but ours use unlabeled data (SET-R) with weakly supervised method to find the parameter setting.

This assessment data is not 5 times in the unlabeled data. Both methods were tested in the same division test in 5-fold cross-validation. All the evaluation results are averaged from 5 folds. For processing the question, use our own statistical POS tagger developed in-house.

B. Experiment Results

1. Comparative Question Identification and Comparator Extraction

Table 1 shows the experimental results. In the table, "unique ID" indicates actions in comparative question identification, "Extraction only" denotes extraction performances comparison when only comparative questions are used as input, and "All" indicates the end to end performance when identifying question results were used in the extraction of the comparator. Note that the results of method J&L's in our collections are very comparable to what is reported in your article.

	Identification only(SET-A+SET-B)			Extraction only(SET-B)		All(SET-B)		
	J&L(CSR)		Our Method	J&L (LSR)	Our Method	J&L		Our method
	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*	0.722	0.833*	0.423	0.479	0.768*

Table 1: performances comparison between our method and Jindal and Bing's method The values with * indicate Statistically significant improvements over J&L (CSR) SVM or J&L (LSR) according to t-test at p<0.01 level.

In terms of accuracy, the J & L's method is competitive to our method of identifying comparative question. However, recovery is significantly lower than ours. In terms of memory, our method outperforms J & L's method by 35% and 22% identity with the comparative question comparator extraction respectively.

In our analysis, the low recovery of the method of J& L's is mainly caused by the low coverage of CSR patterns learned during the test.

In experiments from end to end, our approach performs significantly better weakly monitored the method of J&L's. Our method is about 55% better in the F1-measure. This result also highlights another advantage of our method that identifies and extracts the comparative questions comparators simultaneously using a unique pattern. J&L's method uses two types of standard rules, i.e. the CSR and LSRs. Its yield is reduced due to error propagation. F1-measure method of J&L's "All" is 30%

DOI 10.17148/IJARCCE.2016.5871

and 32% worse than the scores for "unique ID" and "Removal" only respectively, our method shows only small amount of performance decrease (approximately 7-8%).

The effect of the pattern of generalization and specialization were also discussed. Table 2 shows the results. Despite the simplicity of our methods, significantly contribute to the improvement of throughput manner. This result highlights the importance of learning patterns flexibly to capture various expressions of comparative questions. Between 6127 IEP learned in our database, 5,930 generalize patterns, 171 are specialized and only 26 patterns are not generalized and specialized.

	Recall	Precision	F-score
Original Patterns	0.689	0.449	0.544
+Specialized	0.731	0.602	0.665
+Generalized	0.760	0.776	0.768

Table 2: Effect of pattern specialization and generalization in the end to end experiments

To investigate the robustness of our algorithm for different configurations bootstrapping seed yield between two different seed IEP compared. The results are shown in Table 3. As shown in the table, the performance of our algorithm is stable regardless of bootstrapping significantly different number of pairs of seed generated by the two IEP. This result implies that our bootstrapping algorithm is not sensitive to the choice of the IEP.

Seed patterns	# of resulted seed pairs	F-score
<#start nn/$c vs/cc nn/$c?/.#end/	12,914	0.768
<#start which/wdt is/vb better/jjr, nn/$c or /cc nn/$c ?/. #end>	1,478	0.760

Table 3: performance variation over different initial seed IEPs in the end to end experiments

Table 4 also shows the strength of our bootstrapping algorithm. In Table 3, "All" indicates the actions that all pairs of comparing a single seed IEP is used for bootstrapping, and "partial", indicate the performances using only 1,000 pairs random sample of "All". As shown in the table, there is no significant difference in performance.

Furthermore, an error analysis for the cases in which our method cannot extract correct pairs was performed comparing:

- 23.75% of the errors in the extraction of comparison due to the wide selection of patterns our simple strategy of maximum length IEP.
- The remaining 67.63% of the errors come from comparative questions that cannot be covered by the IEP learned.

Set (# of seed pairs)	Recall	Precision	F-score
All(12,914)	0.760	0.774	0.768
Partial(1,000)	0.724	0.763	0.743

Table 4: performance variation over different sizes of seed pairs generated from a single initial seed IEP “<#start nn/%c vs/cc nn/$c?/.#end>”.

2. Examples of Comparator Extraction

By applying our method to bootstrapping the entire data source (questions 60M), 328,364 unique pairs of comparison were drawn from comparative questions 679 909 automatically identified.

Table 5 lists the top 10 states in frequency compared to a target element, such as Chanel, Gap, and our question file. As shown in the table, our method of comparison mining discovers realistic comparators success. For example, for "Chanel", most of the results are the high fashion brands, such as "Dior" or "Louis Vuitton" range while the results of the classification of "Gap" usually contains the Similar clothing brands for young people, as "Old Navy" or "banana Republic". For the basketball player "Kobe", most of the top ranked comparators are also famous basketball players. Some interesting comparators are shown for "Canon" (the name of the company). It is famous for different types of products, such as digital cameras and printers, so you can compare different types of businesses. For example, compared with "HP", "Lexmark", or "Xerox" printer manufacturers, and also compared with the "Nikon", "Sony" or "Kodak", the digital camera manufactures. In addition to the general entities, as a trademark or trade name, our method also found an interesting article on a specific entity comparable experiments. For example, our method recommended "Nikon D40I", "Canon Rebel XTi" "Canon Rebel XT", "Nikon d3000", "Pentax K100D", "Canon EOS 1000D" as terms of comparison for the specific camera product "Nikon 40d ".

	Chanel	Gap	iPod	Kobe	Canon
1	Dior	Old Navy	Zune	Lebron	Nikon
2	Louis Vuitton	American Eagle	mp3 player	Jordan	Sony
3	Coach	Banana Republic	PSP	MJ	Kodak
4	Gucci	Guess by Marciano	Cell phone	Shaq	Panasonic
5	Prada	ACP Ammunition	iPhone	Wade	Casio
6	Lancome	Old Navy brand	Creative Zen	T-mac	Olympus
7	Versace	Hollister	Zen	Lebron James	HP
8	LV	Aeropostal	iPod nano	Nash	Lexmark
9	Mac	American Eagle outfitters	iPod touch	KG	Pentax
10	Doney	Guess	iRiver	Bonds	Xerox

Table 5: Examples of comparators for different entities

Table 6 can show the difference between our mining and query recommendation comparison / article. As shown in the table, usually suggests a mixed set of two types of queries related target entity "Google related search results": (1) specified in sub-queries to the original query (e.g., "Chanel Handbag "" Chanel ") and (2) its comparable entities (e.g.," Dior "" Chanel "). It confirms our claims that mining and query recommendation comparator / item are related but not the same.

Chanel	Gap	iPod	Kobe	Canon
Channel handbag	Gap coupons	iPod nano	Kobe Brayant stats	Canon t2i
Channel sunglass	Gap outlet	iPod touch	Lakers Kobe	Canon printers
Channel earrings	Gap card	iPod best buy	Kobe espn	Canon printer drivers
Channel watches	Gap careers	iTunes	Kobe Dallas Mavericks	Canon downloads
Channel shoes	Gap casting call	Apple	Kobe NBA	Canon copiers
Channel jewelry	Gap adventures	iPod shuffle	Kobe 2009	Canon scanner
Channel clothing	Old navy	iPod support	Kobe san Antonio	Canon lenses
Door	Banana republic	iPod classic	Kobe Brabant 24	Nikon

Table 6: Related queries returned by Google related searches for the same target entities in Table 5 The bold ones indicate overlapped queries to the comparators in Table5.

V. CONCLUSION

We expand a weakly supervised bootstrapping means to identify comparative questions and take out comparable entities simultaneously. In weakly supervised indicative extraction pattern mining method is a pattern-based approach comparable to Jindal and Liu method, but it is different in a lot of aspects such as an alternative of using various class sequential rules and label chronological rules, our process aims to become skilled at sequential prototype which can be able to be used to identify comparative questions and take out comparators simultaneously.

Two important suppositions are designed by using our weakly supervised indicative extraction pattern mining method. This is due to the greatest indicative extraction pattern is probable to be the most exact and persistent pattern for the given query. If a chronological prototype can take out numerous dependable comparable entity pairs, it is very likely to be an indicative extraction pattern.

If a comparable entity pair can take out an IEP, the pair is dependable. Our comparable entity mining outcomes can be used for a business search or product reference system. The results show that our method is capable in both comparative query identification and excavated entities removal. It significantly progress recall in both responsibilities whereas uphold high accuracy.

REFERENCES

[1]. Mary Elaine Califf and Raymond J. Mooney. 1999. Relational learning of pattern-match rules for information extraction. In Proceedings of AAAI'99 /IAAI'99.

[2]. Claire Cardie. 1997. Empirical methods in information extraction. AI magazine, 18:65–79.

[3]. Dan Gusfield. 1997. Algorithms on strings, trees, and sequences: computer science and computational biology. Cambridge University Press, New York, NY, USA

[4]. Taher H. Haveliwala. 2002. Topic-sensitive pagerank. In Proceedings of WWW '02, pages 517–526.

[5]. Glen Jeh and Jennifer Widom. 2003. Scaling personalized web search. In Proceedings of WWW '03, pages 271–279.

[6]. Nitin Jindal and Bing Liu. 2006a. Identifying comparative sentences in text documents. In Proceedings of SIGIR '06, pages 244–251.

[7]. Nitin Jindal and Bing Liu. 2006b. Mining comparative sentences and relations. In Proceedings of AAAI '06.

[8]. Zornitsa Kozareva, Ellen Riloff, and Eduard Hovy. 2008. Semantic class learning from the web with hyponym pattern linkage graphs. In Proceedings of ACL-08: HLT, pages 1048–1056.

[9]. Greg Linden, Brent Smith and Jeremy York. 2003. Amazon.com Recommendations: Item-to-Item Collaborative Filtering. IEEE Internet Computing, pages 76-80.

[10]. Raymond J. Mooney and Razvan Bunescu. 2005. Mining knowledge from text using information extraction. ACM SIGKDD Exploration Newsletter, 7(1):3–10.

[11]. Dragomir Radev, Weiguo Fan, Hong Qi, and Harris Wu and Amardeep Grewal. 2002. Probabilistic question answering on the web. Journal of the American Society for Information Science and Technology, pages 408–419.

[12]. Deepak Ravichandran and Eduard Hovy. 2002. Learning surface text patterns for a question answering system. In Proceedings of ACL '02, pages 41–47.

[13]. Ellen Riloff and Rosie Jones. 1999. Learning dictionaries for information extraction by multi-level bootstrapping. In Proceedings of AAAI '99 /IAAI '99, pages 474–479.

[14]. Ellen Riloff. 1996. Automatically generating extraction patterns from untagged text. In Proceedings of the 13th National Conference on Artificial Intelligence, pages 1044–1049.

[15]. Stephen Soderland. 1999. Learning information extraction rules for semi-structured and free text. Machine Learning, 34(1-3):233–272.

BIOGRAPHIES

Mrs. P. Ragha Vardhani is working as Assistant Professor in the Department of Computer Science and Engineering at RCEW. She has 3.8 years of teaching experience. She has obtained her master's degree in COMPUTER SCIENCE & ENGINEERING from Vaagdevi institute of technology & sciences(JNTUA) in 2012. She has obtained her bachelor's degree in COMPUTER SCIENCE & ENGINEERING from

DOI 10.17148/IJARCCE.2016.5871

Vaagdevi institute of technology & sciences (JNTUA) in 2010. She is a lifetime member of Indian Society for Technical Education (ISTE).

Mrs. Y. Indira Priyadrashini is working as Assistant Professor in the Department of Computer Science and Engineering at RCEW. She has 4.2 years of teaching experience. She has obtained her under graduation degree in computer science in St. Joseph's Degree College, Kurnool (S.K. University). She has obtained her master's degree in Computers Applications, from S.V.U.P.G. Center Kadapa (S.V. University). She has obtained her master's degree in Computer Science and Engineering from G. Pulla Reddy Engineering College, Kurnool. She is a Life time member of Indian Society for Technical Education and Computer Society of India.

DOI 10.17148/IJARCCE.2016.5871