Mining of Comparable Entity with the use of Comparative Question

Deepti A.Barhate1, Prashant Jawalkar 2

1. Deepti A. Barhate Author is currently pursuing M.E (CSE) in BSIOT&R college of Engineering, Wagholi (Pune). e-mail:deepti1.barhate@gmail.com.
2. Prashant Jawalkar, Assistant Professor in BSIOT&R college of Engineering, Wagholi (Pune). e-mail :Prashant.jawalkar@gmail.com

Abtract - Comparison of one thing among things is a usual part of human decision formation process, particularly during an online purchase order. Without comparing it is not fair to buy a product, since it won't give an ideal performance. For instance, if somebody is fascinated in definite products such as cameras, then he /she would need to identify what another possibility are and compare dissimilar cameras before purchasing. This way of comparison action is very common in our day-to-day life but wants high knowledge ability. So, in order to solve this trouble, we are presenting an idyllic way for inevitably mine comparable entities using comparative questions that users dispatched online. It gives a chance to improve the search knowledge by automatically offering comparisons to user. A weakly supervised bootstrapping procedure is employed here for comparative problem identification and comparable entity abstraction by assembling a large online question collection. This result also provides users to enhance new attributes of their interest to the explanation form, so that the next search recovers the provided new attribute information. This technique would overtake the existing system of online shopping.

Keywords— Information extraction (IE), comparable entity mining, part of speech (POS), Information extraction, Robust automated formation of IER.

I. INTRODUCTION

In decision-making process, comparing additional options is one of the necessary steps that we carry out on a regular basis. Though this involves high knowledge skill. For e.g., during online shopping of Computer one must have complete knowledge of its specifications like Processor Speed, Memory, Storage, Graphics, Display, etc. In such case, it becomes problematic for a person with inadequate knowledge to make a good judgement on which computer to purchase and also comparing the different options for the same.

In this paper, our focus is on finding a set of comparable entities provided a user's input entity. For example, provided an entity, Nokia N-95 (cell phone), we want to find comparable entities such as Nokia N82, iPhone, blackberry and so on. To excerpt comparable entities from comparative matter, we should first know whether a question is relative or not.

In the World Wide Web period, a comparison action normally involves search for related web pages covering evidence about the directed products, find contradictory products, read assessments, and classify advantages & disadvantages. In this paper, our focus is on searching a collection of comparable entities specified customer's input entity. For example, assumed an entity, Nokia N-95 (a cell phone), we need to find comparable entities such as Nokia N-82, i-Phone and blackberry etc. To excerpt comparators using comparative matter, we should have to notice whether given question is absolute or not. According to our characterisation, a comparative question needs to be a question with determined to relate at least two entities. Remind that a question covering as a minimum two entities is not a comparative query if it does not need comparison intent. Although, we notice that a query is very likely to be a comparative query if it covers at least two entities. We control this awareness and improve a weakly supervised bootstrapping process to detect comparative queries and abstract comparators instantaneously.

The comparative questions and comparators can be therefore defined by way of:

- **Comparative question**: A question whose goal is to relate two or more objects and it needs to remark these entities clearly in the question.

- **Comparator**: An entity which is a goal of association in a comparative query.

According to the descriptions, Q1 & Q2 further down are not relative questions however Q3 is.
"Noida" and "Hyderabad" are comparators.
Q1. "Which one is better?"
Q2. "Is America the best city?"
Q3. "Which city is better America or Swedan?"
The outcomes will be very suitable in helping users' study of different choices by advising them comparable entities created on other earlier users' needs.

Terms and conceptions:

- **Information Extraction**: The procedure of spontaneously drawing out structured information from unstructured or a semi-structured machine-readable text is named as Information Extraction.

  Approaches for information extraction.

  Comparable entity mining: Comparable entity mining is related with mining the comparable entities from the Text or questions or web mass.

  Sequential Pattern mining: Sequential Pattern mining is mostly related with finding statistically applicable patterns amongst data examples where the values are transported in a sequence.

- **POS Tags (Part-of-speech):** Part-of-speech of a word is a semantic category defined by its syntactic or morphological comportment. Common POS categories are: noun, verb, adverb, adjective, conjunction, preposition, interjection and pronoun. Then there are many classes which arise from different forms of these classes.

II. RELATED WORK

Li et al [1]. proposed a weakly- supervised bootstrapping method to identify comparative questions and extract comparable entities. Author proposed novel weakly supervised method to identify comparative questions and extract comparator pairs simultaneously. We rely on the key insight that a good comparative question identification pattern should extract good comparators, and a good comparator pair should occur in good comparative questions to bootstrap the extraction and

identification process. By leveraging large amount of unlabeled data and the bootstrapping pro- cess with slight supervision to determine four parameters, they found 328,364 unique comparator pairs and 6,869 extraction patterns without the need of creating a set of comparative question indicator keywords. The experimental results show that this method is effective in both comparative question identification and comparator extraction. It significantly improves recall in both tasks while maintains high precision. Their examples show that these comparator pairs reflect what users are really interested in comparing. comparator mining results can be used for a commerce search or product recommendation system. For example, automatic suggestion of comparable entities can assist users in their comparison activities before making their purchase decisions. Also, results can provide useful information to companies which want to identify their competitors.

Author proposed [5] the study of identifying comparative sentences. Such sentences are useful in many applications, e.g., marketing intelligence, product benchmarking, and ecommerce.

Author first analysed different types of comparative sentences from both the linguistic point of view and the practical usage point of view, and showed that existing linguistic studies have some limitations and then made several enhancements. After that they proposed a novel rule mining and machine learning approach to identifying comparative sentences. Empirical evaluation using diverse text data sets showed its effectiveness. An important approach to text mining involves the use of natural-language information extraction. Information ex- traction (IE) distils structured data or knowledge from un- structured text by identifying references to named entities as well as stated relationships between such entities. IE systems can be used to directly extricate abstract knowledge from a text corpus, or to extract concrete data from a set of documents which can then be further analysed with traditional data-mining techniques to discover more general patterns. Author discussed methods and implemented systems for both of these approaches and summarize results on min-ing

real text corpora of biomedical abstracts, job announcements, and product descriptions. Author discussed two approaches to using natural-language information extraction for text mining. First, one can extract general knowledge directly from text. As an example of this approach, they reviewed project which extracted acknowledge base of 6,580 human protein interactions by mining over 750,000 Medline abstracts. Second, one can first extract structured data from text documents or web pages and then apply traditional KDD methods to discover patterns in the extracted data. As an example of this approach, they reviewed work on the Disco TEX system and its application to Amazon book descriptions and computer science job postings and resumes.

Author present a novel approach to weakly supervised semantic class learning[4]. from the web, using a single powerful hyponym pattern combined with graph structures, which capture two properties associated with pattern-based extractions Popularity and productivity. Intuitively, a candidate is popular if it was dis- covered many times by other instances in the hyponym pattern. A candidate is productive if it frequently leads to the discovery of other instances. Together, these two measures capture not only frequency of occurrence, but also cross-checking that the candidate occurs both near the class name and near other class members. They developed two algorithms that begin with just a class name and one seed instance and then automatically generate a ranked list of new class instances. Combining hyponym patterns with pattern linkage graphs is an effective way to produce a highly accurate semantic class learner that requires truly minimal supervision: just the class name and one class member as a seed. Authors results consistently produced high accuracy and for the states and countries categories produced very high recall. The singers and such categories, which are much larger open classes, also achieved high accuracy and generated many instances, but the resulting lists are far from complete. Even on the web, the doubly- anchored hyponym pattern eventually ran out of steam and could not produce more instances. How- ever, all experiments were conducted using just a single hyponym pattern. Other researchers have successfully used sets of hyponym patterns and multiple patterns could be used with our algorithms as well. Incorporating additional hyponym patterns will almost certainly improve cover-age, and could potentially improve the quality of the graphs as well. We present a novel weakly supervised method to identify comparative questions and extract comparator pairs simultaneously. We rely on the key insight that a good comparative question identification pattern should extract good comparators, and a good comparator pair should occur in good comparative questions to bootstrap the extraction and identification process. By Leveraging large amount of unlabeled data and the bootstrapping process with slight supervision to determine four parameters.

III. IMPLEMENTATION DETAILS

In case of determining related items for an entity, our effort is like to the study on recommender structures, which recommend items to a customer. Recommender systems mostly trust on likenesses between items or their numerical associations in customer log data. For example, Amazon commends products to its customers based on their own purchase accounts, similar customers purchase accounts, and likeness between products. However, commending an item is not equal to finding a comparable item. In the example of Amazon, the determination of recommendation is to invite their customers to expand more items in their shopping carts by advising similar or correlated items. In the case of comparison, they help users explore replacements, i.e., helping them make a decision among comparable items. For example, it is sensible to mention iPod speaker or iPod batteries if a user is fascinated with iPod, but we do not compete them with iPod. Though, items which are comparable with iPod such as iPhone or PSP which were got in comparative queries dispatched by customers are challenging to be expected just based on product likeness between them. Though they are all music- players, iPhone is mostly a mobile phone, and PSP is mostly a movable game device. They are same but also dissimilar so appeal comparison with each other. It can be seen that comparator mining and product recommendation are interconnected with each other but not the similar. Our study on comparator mining is associated with the investigation on entity and relative abstraction in information extraction. Bootstrapping approaches have been presented to be very operative in earlier information extraction study. Our work is like to them in case of policy using bootstrapping procedure to excerpt entities with a definite relation. Though, our mission is dissimilar from their in that it involves not only take out entities (comparator extraction) but also confirming that the entities are mined from comparative queries which is usually not essential in IE task.

IV. WEAKLY SUPERVISED PROCESS FOR COMPARATOR MINING

Weakly supervised method is a pattern-based approach similar to JLs method, but it is different in many aspects: Instead of using separate CSR0s and LSR0s, This method aim to learn li marks a token is at the ith position to the left of the pivot and rj marks a token is at j th position to the right of the pivot where i and j are between 1 and 4 in J&L (2006b). Sequential patterns which can be used to identify comparative question and extract comparators simultaneously.

In This approach, a sequential pattern is defined as a sequence S(s1s2 si sn) 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. we will formally define the reliability score of a pattern in the next section. Once a question matches an IEP, it is classified as a comparative question and the token sequences corresponding to the comparator slots in the IEP are extracted as comparators. When a question can match multiple IEP0s, the longest IEP. Therefore, instead of manually creating a list of indicative keywords, we create a set of IEP0s. we will show how to acquire IEPs automatically using a bootstrapping procedure with minimum supervision by taking advantage of a large unlabeled question collection in the following subsections. A. Mining Indicative Extraction Patterns Weakly supervised IEP mining approach is based on two key assumptions: 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 two assumptions, Bootstrapping algorithm as shown in Figure 1 The bootstrapping process starts with a single IEP. From it, we extract a set of initial seed comparator pairs. For each comparator pair, all questions containing the pair are retrieved from a question collection and regarded as comparative questions. From the 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 ones are IEP0s and are added into an IEP repository.

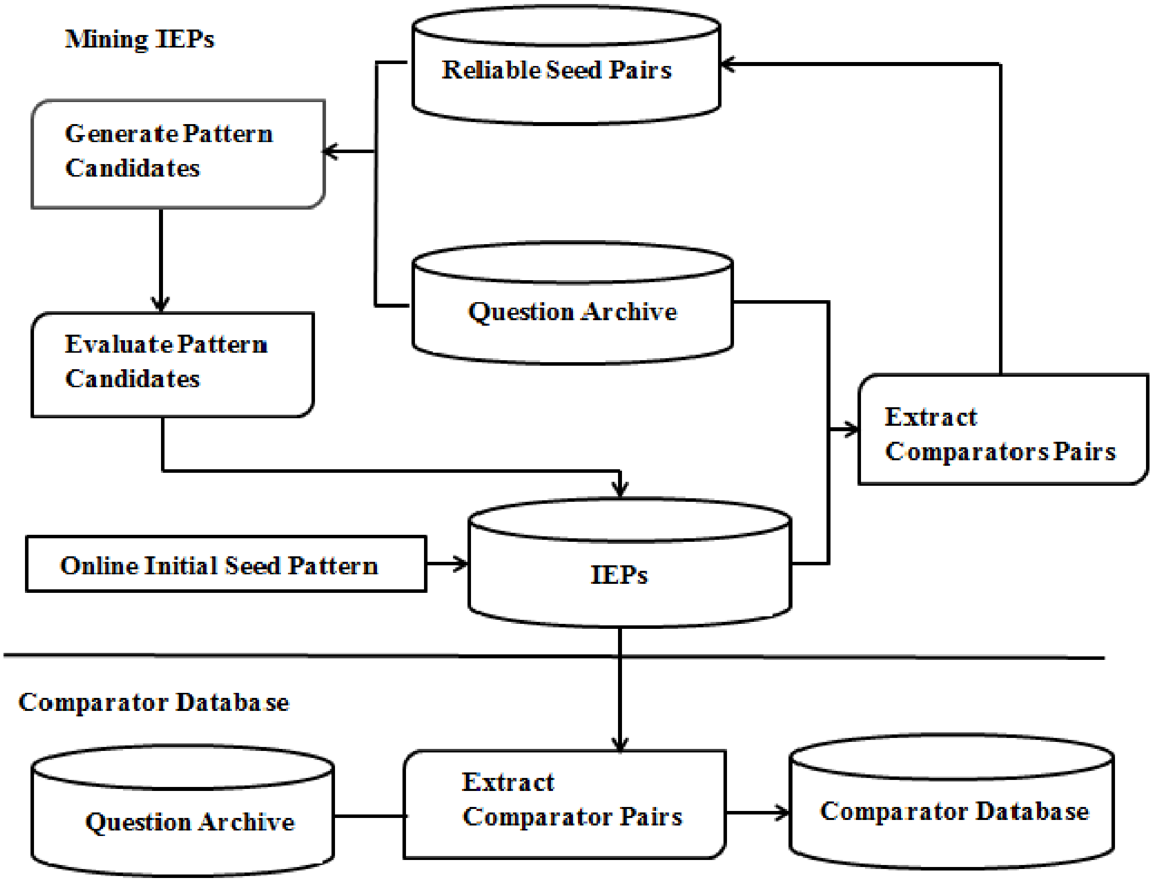

Fig. 1. An Overview of bootstrapping algorithm

Then, new comparator pairs are extracted from the question collection using the latest IEPs. The new comparators are added to a reliable comparator repository and used as new seeds for pattern learning in the next iteration. All questions from which reliable comparators are extracted are removed from the collection to allow finding new patterns efficiently in later iterations. The process iterates until no more new patterns can be found from the question collection.

key steps

There are two key steps in this method:

1) Pattern generation

2) Pattern evaluation

In the following subsections, these steps are explained in details.

Then, the following three kinds of sequential patterns are generated from sequences of questions:

1) Lexical patterns:: Lexical patterns indicate sequential patterns consisting of only words and symbols ($C, #start, and #end). They are generated by suffix tree algorithm (Gusfield, 1997) with two constraints: A pattern should contain more than one $C, and its frequency in collection should be more than an empirically determined number.

2) Generalized patterns:: A lexical pattern can be too specific. Thus, we generalize lexical patterns by replacing one or more words with their POS tags. $2n^1$ generalized patterns can be produced from a lexical pattern containing N words excluding $Cs.

3) Specialized patterns:: In some cases, a pattern can be too general. For example, although a question "ipod or zune?"is comparative, the pattern "<$C or $C>"is too general, and there can be many non-comparative questions matching the pattern, for instance, "true or false?". For this reason,

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?>be produced as a specialized pattern. Generalized patterns are generated from lexical patterns and the specialized patterns are generated from the combined set of generalized patterns and lexical patterns. The final set of candidate patterns is a mixture of lexical patterns, generalized patterns and specialized patterns.

Model

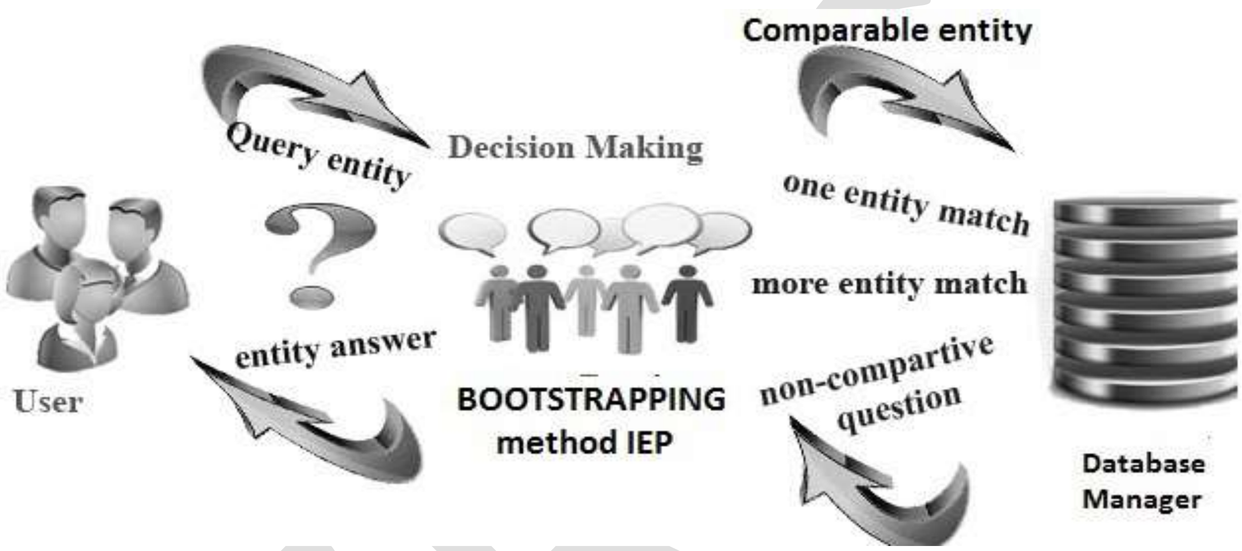

Fig. 2. System architecture of extracting entities

1) **Input :** Given a set of N training examples of the form (x-1, y-1), ..., (x-N ,y-N) such that x-i is the feature vector of the i-th example and y-i is its label(i.e. class), a learning algorithm seeks a function g: X to Y,where X is the input space and Y is the output space. The function g is an element of some space of possible functions G, usually called the hypothesis space. It is sometimes convenient to represent g using a scoring function f: X times Y to Bbb R such that g is defined as returning the y value that gives the highest score:

$$g(x) = \arg -\max-y; (x,y).$$

2) **Process :** Let F denote the space of scoring function. Although G and F can be any space of functions, many learning algorithms are probabilistic models where g takes the form of a conditional probability model

$$g(x) = P(yjx)$$

or f takes the form of a joint probability model

$$f(x,y) = P(x,y).$$

For example, naïve Bayes and linear discriminant analysis are joint probability models, whereas logistic regression is a conditional probability models. There are two basic approaches to choosing org: empirical risk minimization and structural risk minimization. It is assumed that the training set consists of a sample of independent and identically distributed pairs,(x-i, y-i).

3**) Output :** In order to measure how well a function fits the training data, a loss function L. For training example(x-i, y-i),the loss of predicting the value that y is L(y-i,y). The risk R(g) of function g is defined as the expected loss of g.

V. RESULT DISCUSSION

The following Table I shows comparative value for existing system, and Table II shows comparative value for proposed system.

TABLE I. TABLE FOR EXISTING SYSTEM

Chanel	**Gap**	**ipod**
Chanel handbag	Gap coupons	iPod nano
Chanel sunglasl	Gap outlet	iPod touch
Chanel earrings	Gap card	iPod best
Chanewatchesl	Gap careers	iTunes
Chanel shoes	Gap casting	call Apple

TABLE II. TABLE FOR PROPOSED SYSTEM

Mobile	**Television**	**cosmatics**
Micromax	Philips	Oriflame
Samsung	Sony	Yardeley
LG	Panasonic	Amway
Sony	LG	Lakme
HTC	Samsung	Orchid
Lenovo	Toshiba	VOV

GRAPHS:

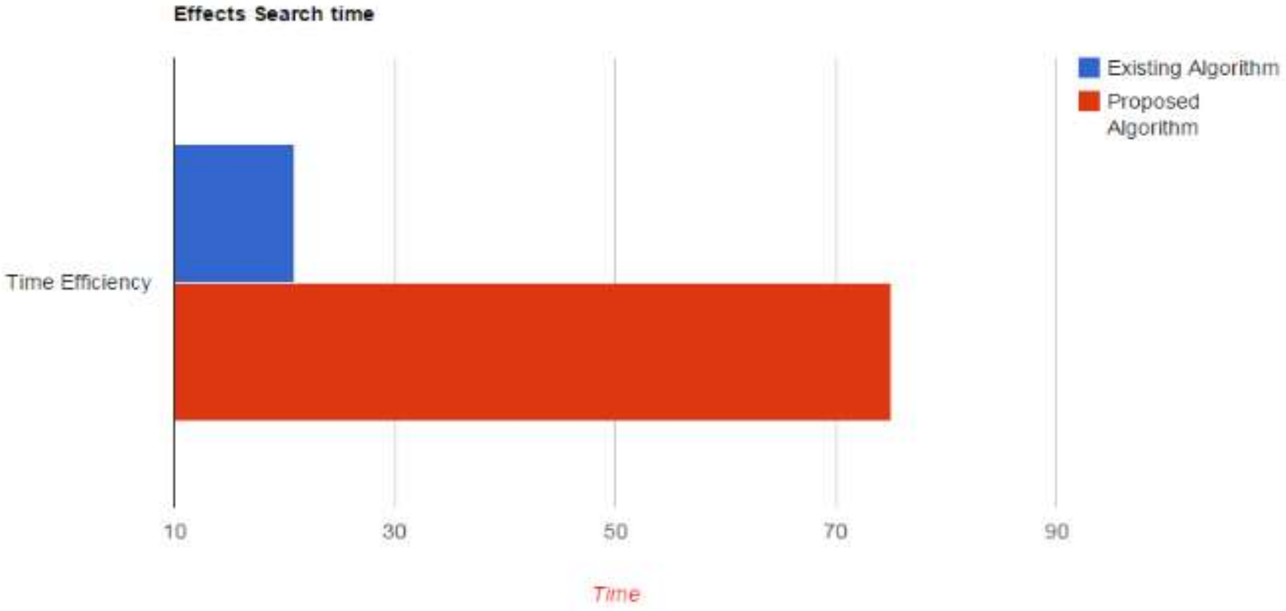

Fig: Time efficiency difference between existing & proposed system

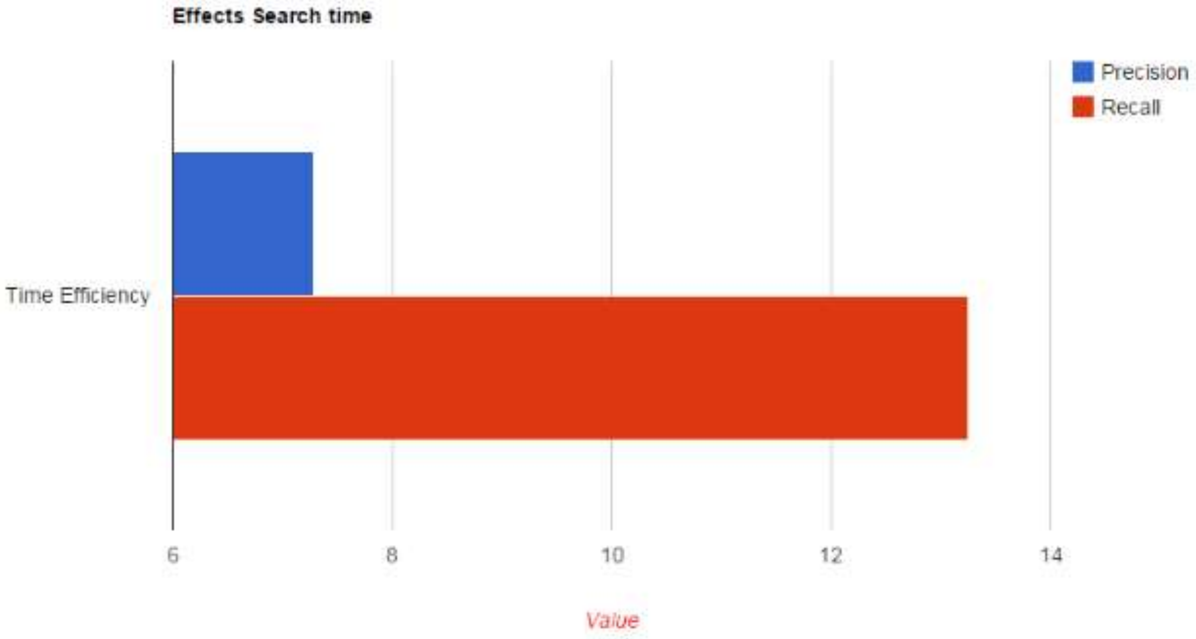

Fig: Precision & recall values

VI ACKNOWLEDGMENT

The authors would like to thank the researchers as well as publishers for making their resources available and teachers for their guidance. We are also thankful to the reviewer for their valuable suggestions..

VII. CONCLUSION

In this paper, we present a new weakly supervised process to recognize comparative questions and excerpt comparator pairs concurrently. We trust on the key insight that a good comparative query recognition pattern should excerpt good comparators, and a good comparator couple should occur in noble comparative queries to bootstrap the extraction and identification process. This technique noticeably develops recall in composed tasks whereas maintain greater precision. Comparator mining conclusion can be useful for commerce exploration or product recommendation association. For instance, automatic suggestion of comparable entities can help out users in their valuation activities earlier than building their acquire judgment. Likewise, the outcome can make available supportive information to corporations which would like to recognize their competitors.

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. 25, no. 7, 2013, 1498-1509*
.

[2] 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.*

[3] Jain, A., and Pennacchiotti, M. 2010, "Open entity extrac-tion from web search query logs". *Proc. 21st Natl Conf. Artificial Intelligence244- 251N.*

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

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

[6] Jindal and B. Liu," Mining Comparative Sentences and Relations" *Proc. 21st Natl Conf. Artificial Intelligence (AAAI 06), 2006.*

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

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

[9] M.E. Califf and R.J. Mooney," Relational Learning of Pattern- Match Rules for Information Extraction "*Proc. 16th Natl Conf. Artificial Intelligence and the 11th Innovative Applications of Artificial Intelligence (AAAI 99/IAAI 99), 1999.*

[10] C. Cardie, "Empirical Methods in Information Extraction Artificial Intelligence Magazine", *vol. 18, 1997, 65-79*