Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 5, May 2015, pg.988 – 993

REVIEW ARTICLE

Comparative Questions Entity Mining: A Review

Vinita A.Pawar[1], B. S. Chordia[2]

[1]Master Student, Computer Science & Engineering SSVPS'S B.S.Deore College of Engineering, India

[2]Assistant Professor Computer Science & Engineering SSVPS'S B.S.Deore College of Engineering, India

[1] vinita.pawar16@gmail.com; [2] chordiabs@yahoo.com

Abstract: Comparative information extraction is one of the hurdles in text mining and an area of research. Comparing two things is a regular habit of human decision making process. Though it is not always known what to evaluate and what are the alternatives as people do comparison of two things for various reasons. Comparative entity mining extracts the comparable or alternative entity from the comparative questions. In this paper, studies the number of techniques that finds or evaluates alternative options and also the techniques that automatically generates patterns for mining.

Keywords— Information extraction, Text mining, Entity mining

I. INTRODUCTION

Text mining is an approach that converts the unstructured text into structured text. Text mining is the method that finds pattern from given text and used to find the new information. Information extraction is the task of filling template information from previously unseen text which belongs to a specific domain. Today information extraction is based on pattern matching.

Almost every day, people are facing the situation that they must have to decide upon one thing or the other. However to have the better decisions, they need to decide one entities from two depend upon their interest. If someone is interested in certain products or services such as laptop or medical treatments then one must know what the alternatives are and how they compare to each other before making a purchase decision. Today is the era of internet and people approaches the internet to get the information so that they can better decision among the interested entities. There are also traditional survey methods which provide information. But better way is to get the information from a web data provided by the search engines. Also searching vast amounts of data is a tedious and wasting of time. Selecting one thing from two comparable things is a very difficult work because a person does the comparison for a dozen of reasons. Tremendous questions are entered by user online which hints what people want to compare e.g. what to buy Samsung or Nokia? Which shows that people are Samsung with Nokia? A questions that compares two or more entities and the entities are mentioned explicitly in the question is a comparative question. The entities which are compared in the comparative question is a comparator. A comparative mining system automatically provides a best option or thing from two or various things from a large quantity of internet data. It is also be very useful in marketing area. Industrialist will know who are the competitive to them. Also customers will easily differentiate the good entity [9].

In the following example q1 & q2 are not comparative questions whereas q3 is comparative question in which "Samsung" and "Nokia" are comparators.

Q1. "Which one is good?"

Q2. "Is Nokia x02 the best mobile?"

Q3. "Which car is better Samsung and Nokia?

II. LITERATURE SURVEY

In the comparable entity mining result there are three categories of analysis is performed first the analysis of recommender system that recommends item to an user, second Jindal and liu's entity mining system, finally different information extraction systems that uses bootstrapping technique for to extract entities with a specific relation.

Information Extraction (IE) is the process of finding the location of certain amount or pieces of data in natural-language documents, then mine structured and meaningful information from unstructured or semi-structured text is called as Information Extraction. In entity recognition information extraction system extraction involves identifying references for particular objects such as names of person, industry, and place. There are two main methods used for information extraction as given below,

1. Rule based Extraction:

In rule based extraction system, pattern based extraction rules are automatically learned for identifying entity or relation of each type. Rapier information extraction system is an example of rule based extraction. Patterns are in the form of regular-expression language; and a bottom-up relational algorithm is used learner is used to generate rules from a labeled corpus examples. Inductive Logic Programming involves the learning of logical rules that are used to identify phrases which are extracted from a document.

2. Pattern based extraction:

In pattern based approaches patterns are created for annotated text fragments (the patterns), where words/phrases are labeled with linguistic information, e.g. POS-tag, syntactic information. The patterns are matched against linguistically annotated text and are used to detect relationships.

Bootstrapping is the methods used for information extraction starts with a small set of given input pattern from a given relation. The extraction system using the input patterns finds seed entries in the plain text and and learns extraction patterns using the seed entries and given text. Extraction patterns are, in turn, applied to text to identify new patterns. In the each iteration, both extracted patterns and identified entries are assigned a confidence score, and patterns and entries with reliable confidence are accepted only. This process continues iteratively until no new patterns are detected.

A. Sentiment Classification

In text and data mining, there is no any work found that directly related to comparative sentences or its identification before Jindal and liu's supervised entity mining. Sentiment classification and opinion extraction are the two only somewhat related to comparative sentences. Opinion texts or sentences are classified as positive or negative classes by sentiment classification. Hearst inspired by cognitive linguistics and classifies entire documents using models. Das and Chen perform stock postings classification using manually crafted lexicon and several scoring methods simultaneously. Tong tracks online discussions about movies over time and generates sentiment timelines that is positive or negative. Turney observes the phrases in the documents and the words "poor" and "excellent" and then apply unsupervised learning technique to their observed mutual information and creates opinions indicative words for classification. Pang performs sentiment classification of movie review by using different supervised machine learning methods. But all their work shows that classifiers work on review is satisfied but on sentences it works poorly as sentence involve very less information. Wiebe works on classification of sentence subjectivity. Adjectives shows negative or positive classification and the method is discovered to find these adjectives. A method is discovered to find adjectives that are indicative of positive or negative opinions. A similar method for finding noun is also proposed. To analyse opinion in customer review, various supervised and unsupervised techniques are discovered. But they only identify features of product that was posted by customer and classify the negative or positive opinions [11] [12].

B. Recommender System

The research on recommendation system is similar to the method of finding comparative items for a user input entity which will suggest items to a user input. The similarities among the items, their statistical correlations in user log data are the two important things that recommender system and their techniques follow. For example, online seller helps by suggesting products to its customers based on their earlier purchase record of customers, and similarity among products. However recommending and finding a comparable item are two different things. In Amazon recommendation system the aim of recommendation is to make ready their customer about their product in such a way that the customers will add more items to their shopping bag by providing related or identical entity. But the comparative system explores alternatives so that user can make a decision among alternate items. Consider the example, it is reasonable to recommend "Samsung speaker" or "Samsung batteries" if a user is interested in "Samsung cell," but we would not compare them with "Samsung cell." However, items that are comparable with "Samsung" mobile such as "Nokia" or "Spice" mobile that were found in comparative questions provided by users are difficult to be detect simply based on common quality between them. But these are all music players and accessories; "Samsung" is mainly a mobile phone. They are similar but also different

therefore have comparisons with each other. Therefore the mining of comparator entity and item recommendation are somewhat related but not equivalent [3].

C. Supervised Entity Mining

Jindal and Liu have done the works on mining comparative sentences and relations using class sequential and label sequential rules. In this method class sequential rules and label sequential rules trained from the labeled corpus for identification of comparative sentences and extraction of comparative relations. CSR is a classification rule and LSR is labeling rule. The class sequential rule maps the sequence pattern S to a class say C and label rule convert sequence pattern which is input to a sequence of label by replacing one token from sequence with a required label. The anchor is extracted if and only if its respective label in the labeled sequence is the extracted comparator.

Supervised entity mining method treated identification of comparative sentence as a classification and relation extraction as an information extraction problem. First a list of indicators was manually created that are likely indicators of comparative sentences. The list includes morphemes such as more, *–er*, less *-est* and also many other indicative words for comparisons, e.g., *beat*. They compiled a list of 79 such keywords. Four exceptions are there *most, more, least and less* are considered as individual keywords because their use are varied and relates to individual patterns. Totally, there were 83 keywords and helps in generate part-of-speech date sequence. A manually annotated corpus with comparative or noncomparative class was used to create sequences and class sequential rules were mined. A Naive Bays classifier was trained using the class sequential rules as features and then used to identify comparative sentences. Two comparators are labeled with $ES1 and $ES2, the feature is label with $FT in every sentence. The method was only used with noun and pronoun. The fourth label $NEF, i.e., nonentity feature is used to differentiate those noun or pronoun that are not comparators or features. All labels were used together with tokens li & rj1 (tokens at specific positions), #start indicates the start of sentence or question, and #end indicates the end of sentence or question, to generate sequence data. The sequences having only single label and low support greater than 1 percent is consider, and then label rules were generated. When applying the learned label sequential rules for extraction, label sequential rules with higher confidence were applied first [1] [2].

Several disadvantages are there in the Supervised Entity Mining experimental setups. The keyword list is manually created and there is no rule to select the keywords to be included in the list. The whole performance of the system depends on these keywords. And also there is no guarantee that list is complete. Comparative questions and sentences are asked by the user in different ways. And a large tagged training text is required to achieve high recall and it is very costly. The rule used in this method is mainly the combination of keywords and POS tag and result in low recall. Two rules are used to identify the comparative sentence [1] [2].

D. Comparative Type Classification

Seon Yang and Youngjoong Ko uses comparative type classification to extract comparative entities and predicates from texts to build a Korean comparison mining system In first task comparatives are extracted from each sentences of documents and *comparatives* and *non-comparatives* are two categories in which sentences are divide. For this they construct linguistic based keyword set K_{ling} set but it is not sufficient to collect all the actual comparison words or expressions. Hence they construct comparison lexicon which is the union of K_{ling} and other keywords for comparative expressions. Each element of comparison lexicon is a CK element. Next they classify the comparative sentences into seven types' greater or lesser, equality, superlative, similarity, pseudo, difference, implicit comparison and then employ transformed base learning for comparative sentence classification. In second task they mine comparative thing and predicates by considering the properties of each type. Consider the, sentence "President is bigger than Prime Minister" belonging to superlative type. Here "president" as a subject entity, "prime minister" as an object entity, and bigger is a comparative predicate. For mining comparative entities and predicates they use POS tag and then generates pattern [6].

E. Bootstrapping

Jindal and Liu's supervised entity mining methods mines the comparative sentences using rules and indicative keywords. Bootstrapping is very important and useful method in earlier information extraction research. The similarity work is bootstrapping technique used by earlier research. The bootstrapping generates extraction pattern to extract entities with a specific relation. But difference is that in previous work only entities are extracted and it is not confirmed that they are extracted from comparative sentence or which is generally not important in information extraction. Some information extraction systems are studied that uses bootstrapping approach that creates extraction of patterns.

Ravichandran and Hovy works on how generation of surface text patterns automatically and used them in finding the answer of a question in a question answering system using machine learning. A large tagged corpus is built from the Internet using a learning technique of bootstrapping by providing a few examples of each type of question to AltaVista. The documents are returned from search engine and patterns are generated automatically and standardized. These patterns are applied to find answer to new questions. This method works well for questions like birthdates or location and not for questions of the type definition. Also no external knowledge can be added to these patterns [4].

Ellen Riloff developed a system called AutoSlog System. AutoSlog learning of text extraction rules from training examples. For extracting information from text domain-specific dictionary of concept nodes is used. A concept node is a rule which includes a semantic constraint and a word "trigger". If the text contains the trigger and system finds it then node's conditions are satisfied, activates concept node and the node definition is extracted from the context. The system identifies a sentence annotated with a slot filler and semantic tag. Then, it checks its list of heuristics and sees if the part slot filler in sentence containing match any of the heuristics. Each heuristic covers only a one slot extraction. Learning of extraction rule involves uses of semantic tagger and constraint. It handles only free text and similar concept nodes are not merged. In AutoSlog extraction patterns are automatically created using heuristic rules as shown example in table 1 [7].

Table 1: AutoSlog Heuristic [5]

PATTERN	EXAMPLE
<sub> passive-verb	<victim> was murdered
<sub> active-verb	<prep> bombed
active-verb <dobj>	bombed <target >
passive-verb <dobj>'	killed <victim>
noun prep <np>	bomb against <target>
passive-verb prep <np>	was aimed at <target>

Ellen Riloff (1996) developed a system called AutoSlog-TS**.** This system creates extraction patterns in the form of dictionaries from unlabelled text. AutoSlog-TS is similar as AutoSlog with some extension. AutoSlog-TS need only a corpus pre-classified with respect to each document's relevance to the task of interest. Heuristics are used to create extraction patterns and is only for all noun phrase present in the texts. A major extension is that it allows more than one rules to apply if more than one matches the text and thus more extraction patterns are created. Statistics data will confirm which pattern is needed to be best for the domain. In a second stage evaluation of pattern takes place. The corpus is processed twice and relevance for each pattern is calculated and uses them to rank the extraction patterns. A potential problem with AutoSlog-TS is that there are undoubtedly many useful patterns buried deep in the ranked list, which cumulatively could have a substantial impact on performance. The current ranking scheme is biased towards encouraging high frequency patterns to float to the top, but a better ranking scheme might be able to balance these two needs more effectively [5].

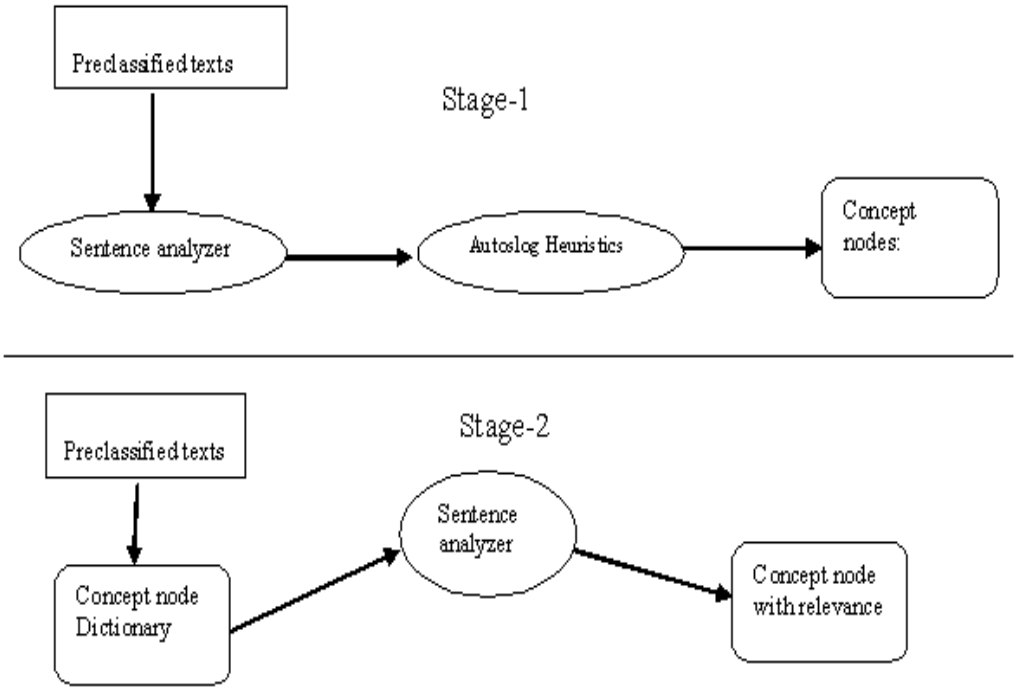

Figure 1: Auto Slog –TS Flowchart [5]

The RAPIER system makes the use of documents samples and filled templates to generate pattern-match rules. By using rules system directly finds and perform the operation of extraction of fillers of given slots in the template. The system uses bottom-up learning algorithm to find fillers in the slots and this limits search without considering constants limits, and to allow high precision by suggesting more specific rules. To achieve the beam search randomly selects two rules to find the best generalization of the pair of rules, taking a least general generalization. Then adds the constraints still the proposed rule work fine on the training data. RAPIER system can allow on semi-structured text a single-slot extraction to be performed [13].

III. COMPARISONS

Title	**Methods/Approaches**	**Advantages**	**Disadvantages**
Mining Comparative Sentences and Relations	Identify or detects comparative sentences from the data and also extract relations indicating comparison	Finds an entity is to directly compare it with a similar entity.	Gains high precision but gives low recall.
Identifying Comparative Sentences in Text Documents.	Supervised Entity Mining: sequential rule (CSR), label sequential rule(LSR)mining and machine learning combination	Extract comparative sentences from text is useful in many areas	Gains high precision but gets from low recall.
Mining Knowledge from Text Using Information Extraction.	Information extraction extracts knowledge or structured data from not structured text	Information Extraction is extracting structured data from semi-structured unstructured documents of web	Building corpus is the requirement of information system and is not reduced or eliminated.
Relational Learning of Pattern Match Rules for Information Extraction	Required data can be extracted through natural language texts	Involves research on relation and entity mining in information extraction	Use of limited syntactic and semantic information by patterns to identify slot fillers and near text.
Learning Surface Text Patterns for a Question Answering System	Learns regular expressions from the internet automatically, for given types of questions	This method works well for questions like birthdates or location	System does not make distinction between upper and lower case letters. Not suitable for definition type question. No external knowledge can be added to the patterns

Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping	AutoSlog-TS : creates dictionaries of extraction patterns from unlabelled text	Allows multiple rules to apply if one or more matches the context and generates multiple extraction patterns.	There are undoubtedly many useful patterns buried deep in the ranked list, which cumulatively could have a substantial impact on performance.
Automatically constructing a dictionary for information extraction tasks	AutoSlog: involves learning of text extraction rules from training examples	For extracting information from text domain-specific dictionary of concept nodes is used. Patterns are automatically generated using concept node.	It handles only free text and similar concept nodes are not merged

IV. CONCLUSIONS

In the writing discovering methods we move towards the effort to differentiate recommendation system, and entity mining, which are related but not same. Recommender systems recommend items to a user and entity mining, explore the alternatives to user to make the alternative decision. Supervised entity mining method mines the comparative sentences and relation using sequential rule in addition to manually created indicative keyword which is a time consuming which result in low recall. Also study various information extraction systems that use bootstrapping method for pattern generation that helps in extracting entities with their specific relation. They have their own advantages and disadvantages.

REFERENCES

[1] N. Jindal and B. Liu, "“Mining Comparative Sentences and Relations," in *Proc. 21st Nat’l Conf. Artificial Intelligence (AAAI ’06),* 2006.

[2] N. Jindal and B. Liu, "Identifying Comparative Sentences in Text," in *Int’l ACM SIGIR Conf. Research and Development in Information Retrieval*, 2006, pp. 244-251.

[3] B. Smith, and J. York G. Linden, "Amazon.com Recommendations: Item-to-Item Collaborative Filtering," *IEEE INTERNET COMPUTING*, vol. 7, pp. 76-80, Jan/Feb 2003.

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

[5] E. Riloff, "Automatically Generating Extraction Patterns from untagged text," in *Proc. 13th Nat’l Conf. Artificial Intelligence,* 1996, pp. 1044-1049.

[6] Seon Yang and Youngjoong Ko, "Extracting Comparative Entities and Predicates from Texts Using".

[7] E. Riloff, "Automatically constructing a dicti onary for information extraction tasks," in *In Proc. of the 11th National conference on Artificial Intelligence*, 1993, pp. 811–816.

[8] E. Riloff and R. Jones, "Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping," in *Proc. 16th Nat’l Conf. Artificial Intelligence and the 11th Innovative Applications of Artificial Intelligence Conf*, 1999, pp. 474-479.

[9] C.-Y. Lin, Y.-I. Song, and Z. Li S. Li, "“Comparable Entity Mining from Comparative Questions," in *Proc. 48th Ann. Meeting of the Assoc. for Computational Linguistics*, 2013.

[10] S. B. Huffman,.*Learning information extraction patterns from examples*.: In Lecture Notes in Computer Science. Connectionist, Statistical, and Symbolic Approaches to, 1996, pp. 246-260.

[11] C Kennedy, *Comparatives, semantics of. In Enclycopedia of Language and Linguistics*, 2nd ed. Elsevier, 2005.

[12] F. Moltmann, "Coordination and comparatives," MIT, Cambridge, Ph.D dissertation 1987.

[13] M.E. Califf and R.J. Mooney, "Relational Learning of Pattern-Match Rules for Information Extraction," in *Proc. 16th Nat’l Conf.Artificial Intelligence and the 11th Innovative Applications of Artificial Intelligence*, 1999.

