Effective Information Retrieval Algorithm for Electronic Market Goods Selection

Mohammad Ali Tabarzad and Caro Lucas

***Abstract*—** One of the issues that researchers are interested in electronic markets is how to describe and delivery of goods. To complete these issues there are so many offers each with specific benefits and defects. One of these methods is that besides the properties of goods usually available in goods' description, there is another group of properties as descriptive properties and other properties which have not been resided in any other property categories will be set in this category beside each other, and will be searched like a text using available information retrieval algorithms. In this paper an appropriate method for searching these properties will be offered based on ngram algorithms and search results will be compare users scoring results.

***Keywords*—**electronic market, multi attribute goods, ngram, information retrieval.

I. Introduction

One of the issues that researchers are interested in electronic markets is how to describe and delivery of goods. To complete these issues there are so many offers each with specific benefits and defects. Some researchers tried using predefined descriptions for goods [1]. In this method which is one of the most usual available methods for goods description problems like searching and finding proper good can be handled easily and fast [2],[3]. The thing that we have to consider in these methods is a skillful person with high ability of description and fluency which can define which attributes should be set in good description and which not. Also there are some problems in decision making about the issues which are not very important or are similar in most of the good models or is true about some models.

Another considered way by researchers is using a text for goods description in natural language and using description logic beside that for extracting available meanings in the text [4]. The problem about this issue is considering new and inadequate meaning extraction from text that still needs wide spreading researches itself, using these methods is not a suitable solution for this problem. From the other hand it is possible that adding new meanings to old ones previously added may encounter confliction, unless the system has been designed totally exact and complete. This issue makes these kinds of systems' development too hard.

The other way which was mentioned in [5] is based on the issue that besides the description of goods available in first method for goods' description, there would be another group of properties as descriptive properties and other properties which have not been mentioned in any other property categories would be set in this category beside each other, and the would be searched like a text using available information retrieval algorithms. In this passage an appropriate method for searching these properties will be offered based on ngram algorithms and search results will be compare users scoring results using this algorithm. Before starting discussing proposed method for market's goods searching, first usual text search methods and then their benefits and defects for this problem should be discussed. Two general categories of these methods are perused as below.

II. Single word algotirhms

In these ways any word separated by space or separator characters are used as usable elements in these category of algorithms and these similar words will be scored by available different weighting methods. More complete explanation about these ways is available in [6].

The main problem of this way is that these kinds of methods do not consider mistyping and treats mistyped words as a new word that is not like previous word and actually is not calculated in scoring. However there is no problem about ordinary texts in which important words are repeated some time, because in case of omitting one of these words, just respective text's weight would be some deal inconspicuous, but as mentioned before in written texts in descriptive properties in which a meaning is just repeated once, losing one word may lead to omission of respective good from search list.

The other problem which make usage of single way algorithms hard, is that in most of these algorithms gaining a desired result needs knowing respective language grammar very well, and there are issues like stemming for omission of general prefixes and suffixes and proper knowledge of language grammar for specifying prefixes and suffixes. If we

This work is supported in part by grants from TAKFA (National Information and Communication Technology Agenda, High Council of Informatics, Iran).

M. A. Tabarzad is with the Center of Excellence: Control and Intelligent Processing, University of Tehran, Tehran, Iran (phone: +98-021-88027757; fax: +98-021-88633029; e-mail: m.tabarzad@ ece.ut.ac.ir).

C. Lucas is with the Center of Excellence: Control and Intelligent Processing, University of Tehran, Tehran, Iran and School of Cognitive Sciences, IPM, Iran(e-mail: lucas@ipm.ir).

want that discussed market can support goods with available descriptions in different languages and also available good be searchable in different languages, completely knowing all those languages is inevitable and this can lead to more complexity of market structure and requirement of continuous updating of market's search core.

III. NGRAM WAY

In this way n is the key name defines the size of window passing over text. Words available in passage are converted to the words with fixed length (n): A slipper window with length n passes over words and any part of word which settles in slipper window is added to created text from main text as a new word. After this procedure a new text with number of words more than in first text and with length n will be created that will be used for those general algorithms in single word way. More complete information is available in [6].

The advantage of these ways is that usual mistyping's and prefixes and suffixes will not omit a word from search list any more, also there would be no need for stemming for different languages. Defect of this way for using in market is that regarding to less number of words and meanings in description properties, available prefixes and suffixes can be scored more than real words which results to increase of wrong retrievals issues.

To solve these problems proposed way creates words using primary concepts of ngram and some deal solve the mistyping and stemming issues and then by use of a weighting method based on expression not words solves the improper weights problem.

IV. PROPOSED WAY TO SEARCH DESCRIPTION OF GOODS

In case of searching about description of goods, and if price property of new buy request matches good's price, similar words or expression would be searched in buy request description and good's description and number of similar cases would be considers as a scoring factor. Main difference of these descriptions and ordinary searchable texts with information retrieval algorithms is these texts are very small in spite of ordinary texts. And also word by word are very important in these texts and mistyping or using as plural or singular in compared texts can not be ignored.

Another issue is that meanings and words repeats in ordinary texts for example an important factor in most information retrieval algorithms is using repetitive number of words in text as a factor of that text with respective query. But in description properties usually repetition of a word or a meaning is just one time and as s result these differences represents lack of a new way for more appropriate retrievals from these texts.

Use

A. *Proposed way to retrieve description of goods*

In this way just like ngram way a window with length of n is used in order to making words in new passage. After making each new word three different values with this formula $t = (d_t, w_t, g_t)$ will be kept in created index for that word about good's description. Value d_t is the number of good that this description refers to that and actually shows which document is this word related to. Then markets the number of the primary word which new words are made over it. This number that is w_t shows the order of word in primary text. g_t is number of words which have been made from that word. E.g. if "spinning rims" was available in respective description and if using 2gram n=2, description will be converted to the words " sp,pi,in,nn,ni,in,ng,,g_ ri,im,ms,s_". Now for example for the first word made from "rims" that is t="ri' d_t is the number of t document among documents and w_t =2 and g_t =1 will be set in index. After all created words were set in the index, all these steps will be repeated for searching expression too which can be user search. After specifying index in searching expression it's expression scoring turn.

B. *Way to give points to primitive words*

First words (which is this way all created words have the same w_t) should be scored so that in case of mistyping or using prefix or suffix in word takes some of the total score and by this way words with near stems will be pointed certainly.

Method of scoring is so that for each two sub-word obtained from main word in searching expression and two equivalent sub-word with equal distance in available description in index, If so, one point will be added to score obtained by that word. E.g. if there is $q_n = (w_{q_n}, g_{q_n})$ for word q_n created from searching expression in index and also there is $t_m = (d_{t_m}, w_{t_m}, g_{t_m})$ available in index for word t_m from description expression for all four words $q_n, q_{n'}, t_m, t_{m'}$ one point will be scored considering following conditions:

1) $n > n', m > m', w_{t_m} = w_{t_{m'}}, w_{q_n} = w_{q_{n'}}, d_{t_m} = d_{t_{m'}}$
2) $q_n = t_m, q_{n'} = t_{m'}$
3) $g_{t_m} - g_{t_{m'}} = g_{q_n} - g_{q_{n'}}$

If two primitive words are completely equal, they will gain points for all binary combination. If there is one mistype error in end of the word for example, will lose score for those binary combinations which one part contains mistyped error only, and binary combinations of first of the word will be scored. Also if searching word does not contain plural sign but available description word is in plural format, available word in description will gain the maximum score as each binary combination of searching word is scored. In this way two mentioned problem in usual ways are some deal solved.

C. *Normalizing the points of primitive words*

In proposed way words with longer length will be scored for

each different binary combination, because in 2gram way for example, which is used here for making secondary words, and using followed way for scoring, each word with length n can be maximum scored for its $\frac{n(n-1)}{2}$ possible binary combinations. To avoid this problem final score is divided by this figure so that all words have a score between zero and one (depending on similarity of two words). After scoring summation of all words' scores will be divided by count of adaptable words.

D. Way of giving points to expressions

Though proposed way has benefits compared to other ways, there is a defect that some improper words in meaning may be scored possibly so if one part of two words were like each other, the algorithm specifies them as stem-mate and gives them even a small score. In order to decrease any possible mistakes in scoring that may be encountered as this case's result it should be considered that usually if searching words which are set near each other, were scored, and some words in searching description are scored for these words, because there is meaning relationship between searching words, in case of correct specification, there should be meaning relationships between two description's available words. So it is necessary that the scored words in were near each other description.

We can use this issue so that if the scored words in searching description are near each other, they are probably correct combinations that a high score should be considered for them then if two words $w_q, w_{q'}$ in first stage scores $w_t, w_{t'}$ in description expression and if $|w_t - w_{t'}| \leq |w_q - w_{q'}| + k$ (in which k is maximum allowed distance between two words) one point is added to score part of the expression.

Now using this if two words not relative in meaning but have achieved points for their words, as they are really improper in meaning, they are possibly located in a far distance from each other that this causes them to lose the second stage's points. Finally in this stage in order to avoid dependency of scoring to number of available words in expression and increasing words doesn't lead to expression increase, like previous stage achieved score should be divided by $\frac{n(n-1)}{2}$ that is the maximum score of binary combinations (in which n is the number of available words in expression). Total score is weight average of word parts and expression part's scores that expression part's weight should be considered multiple times rather that word part because of its more importance.

Cause of why increasing points in second stage is constant and not relative to achieved weight in first stage is that the algorithm may specify a word for its stem similar to searching word that has been came after first three words accidentally. In this case it seems that this combination can be as related to search as two words have been completely specified similar or have been came respectively because it seems that words' stems should have a logical relation to each other. On the other side words which have been specified wrongly similar to a word of searching in stem, as they are really not relative, another word which is again specified wrongly in stem can not be found until far distance possibly and so this combination don't achieve any points in this stage and will be in a low score in total scoring.

V. Experiment Results

For testing proposed description property scoring method in store, a real store is needed in which deals have been held there for a specific time. To do so some of a car available in a local newspaper from three days containing 400 items were selected that 100 of them where used in this test randomly. In first stage 10 questionnaire containing 10 cars were prepared and each one was given to a user.

Users had to score properties' importance extracted for cars from one to five respectively with values "very least importance, least importance, ordinary, important and very important" and then select their car color priorities among available colors for cars in questionnaire. Then users had to score 10 goods in available questionnaire from first to tenth order by their desires. Then search system scored goods from first rank to tenth considering properties importance for users and user priority interests for available colors and finally two different scores were obtained by users and system that will be compared as follow.

The first factor that can simply define similarity of system's scoring to user's scoring is that difference summation average of user's scoring with that good's equivalent scoring by system will be considered as a factor for similarity amount of these two scorings. In this method in ideal case that system and users scores exactly the same and in worst case that users scoring is from 1 to 10 and system scores goods as 10 to 1 a difference score of 50 will be made. In other cases score difference is a number between 1 and 50 and so using this formula $100 - 2x$ in which x represents the score difference as similarity percent of two ways.

In last test scoring difference average was 8.4 that represent 83.2 % similarity of system scoring to users scoring referring to formula. Also system scoring's average difference for each rank was calculated from its real score which had the maximum difference of 1 and 3 for ninth rank and minimum of 0.5 for third, fifth and sixth ranks. Score difference is demonstrated totally in Fig. 1.

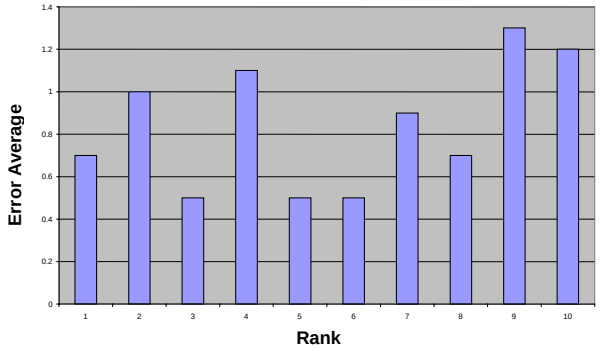

Fig. 1 Error Average in rank scorings by search system

Also standard deviation of each scoring proportional to real calculated score has been calculated for referred scorings which has the maximum of its value in fifth rank equal to 0.75 and the minimum value of 0.35 in first and third ranks. Fig. 2 demonstrates the standard deviation for all ranks.

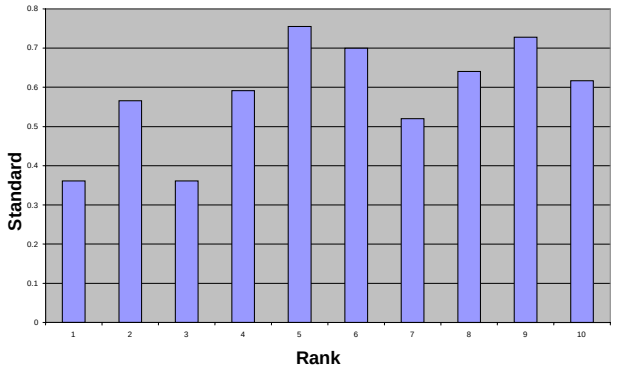

Fig. 2 Standard deviation in rank scoring by search system

According to results from the experiments and watching the high percentage of similarity of system scoring with user scoring it seems that system performance is suitable. Also considering that complete primary knowledge of requested good is no more necessary for description search part; there is no need to completely describe goods in store. Also considering this information retrieval way's benefits that acts very better in small texts compared to usual information retrieval ways, using that seems suitable in buying and selling goods environment.

VI. Conclusion

In this article a new way to offer goods in store was discussed that for description property search a new algorithms based on ngram information retrieval was presented. In continue different issues for using this algorithm in research fields that deals with small amount of data or non frequentative data can be discussed and compared with general algorithms in these special environments. Also weighting measure of this model can be set for different models that both of these issues can prepare appropriate future research fields.

References

[1] E. Fink, J. Johnson and J. Hershberger, "Fast-paced trading of multi-attribute goods," *In Proc. of the IEEE International Conference on Systems, Man and Cybernetics*, vol. 5, pp. 4280-4287, 2003.

[2] E. Fink, J. Gong and J. Hershberger, "Multi-attribute exchange market: Search for optimal matches," *In Proc. of the IEEE International Conference on Systems, Man and Cybernetics*, vol. 5, pp. 4140-4146, 2004.

[3] E. Fink, J. Johnson and J. Hu, "Exchange market for complex goods: Theory and experiments," *Springer, Netnomics*, vol. 1, pp. 21-42, 2004.

[4] T. Di Noia, E. Di Sciascio, F. M. Donini and Marina Mongiello, "A System for Principled Matchmaking in an Electronic Marketplace," *International Journal of Electronic Commerce*, vol. 8, no. 4, pp. 9-37, 2004.

[5] M. A. Tabarzad, C. Lucas, and N. Jafarzadeh Eslami, "A New Method for Complex Goods Selection in Electronic Markets," *Transactions on Engineering, Computing and Technology*, vol. 14, pp. 105-110, 2006.

[6] R. Baeza, Y. Berthier, R. Neto, "Modern Information Retrieval," *ACM Press NY, Addison-Wesley Harlow, England*, 2000.