International Journal of Psychosocial Rehabilitation, Vol. 24, Issue 05, 2020
ISSN: 1475-7192

Pragmatic Security for Efficient Web Search Using Amalgamation

T.P. Anithaashri, M. Saranya and G. Ravichandran

Abstract--- *Natural language processing alone does not improve the accuracy of information retrieval systems. If the focus is shifted to short phrases rather than full documents, the situation becomes extremely different. The proposed system uses the combination of game theory, natural language processing technique and steganography to obtain high accuracy retrieval of images which have been annotated with a descriptive textual caption. The natural language techniques also allow additional contextual information to be derived from the relation between the query and the caption along with high security to store the information in a database, which can help users to understand the overall protected collection of retrieval results.*

Keywords--- *Natural Language Processing, Phrase Matching, Steganography, Game Theory.*

I. INTRODUCTION

Text information retrieval is concerned with finding documents which match against user's query, and assigning a measure according to the closeness of the match. Natural language processing (NLP) can provide rich information about the text, and it might appear reasonable that this would result in better retrieval than conventional "bag of words" approaches. Experiments were done in which simple keywords were augmented with compound terms consisting of pairs of keywords. This paper will look at a technique called phrase matching, which attempts to use lightweight, symbolic natural language analysis to improve retrieval accuracy. It relies on looking for combinations of words which stand in certain modification relationships. Firstly, it recursively explores the structure of the caption and query, checking that terms stand in equivalent modification relations in the two phrases. This also allows the match score to be finely tuned and special cases such as negation to be handled. Secondly, by means of a further algorithm called context extraction, information about non-matching parts of the caption, related to the parts which did match, can be obtained. . This is an important step, because it provides information which is unavailable without natural language analysis, and shows that NLP can contribute in adding new functionality as well as improving accuracy. In section 2 of this paper, the phrase matching algorithm have been explained and give some evaluation results. Section 3 explains on to context extraction. Some conclusions and suggestions for future work are presented in section 4.The title or phrase or words in images based on which the web search is done can be made secure and invisible by embedding them using steganography and game theory.

II. RELATED WORK

The syntactical search has been used in this paper for to find the finite set of parser with the dependency structure and make the search in a dependent manner [2]. The information retrieval has been done exactly by

T.P. Anithaashri, Associate Professor, Department of Innovative Informatics, Institute of CSE, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Thandalam, Chennai. E-mail: shri3krra@gmail.com

M. Saranya, Assistant Professor, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Thandalam, Chennai. E-mail: saranyam.sse@saveetha.com

G. Ravichandran, Research Scholar, AMET University. E-mail: vpigcet@gmail.com

DOI: 10.37200/IJPR/V24I5/PR2020129
Received: 07 Mar 2020 | Revised: 26 Mar 2020 | Accepted: 04 Apr 2020

matching terms in documents with those of a query[13]. However, these lexical matching methods are imprecise when employed to match a user's query[14][15]. This emanates from the fact that there are generally many ways to express a given concept, by use of synonyms. In some circumstances, the literal terms in a user's query may not match those of a relevant document[8],[9][10][22][23].

III. PROPOSED SYSTEM

3.1 The INVIP System, Reconnaissance and Steganography

INVIP (Identical Natural Language Visual Information Pointer) is a retrieval system for databases of digital photographs, intended for operation over the World Wide Web. The photographs are annotated with captions, typically between 05 and 25 words in length, which describe the subject matter of the image through the reconnaissance (RECCE) part of game theory. The reconnaissance is used to obtain information about the users strategic targets. The advisability of reconnaissance before retrieval can be investigated by considering the problem as a game of strategy. Let us consider two strategies namely, expected outcome and unexpected outcome from the retrieval of information. Assuming that the expected outcome, requires to retrieve all the information. For simplicity, let us assume that the expected outcome has two courses of action, namely retrieval with the concern matches and leaving the other matches merely. A strategy for the retrieval of expected outcome will be a set of instructions which directs the search to infer the information that it requires. Thus by using reconnaissance the information relevant to the expected outcome as a result of search can be obtained. This step helps to fix the strategy according to the keyword used to search the information.

The system is intended for casual users, and it is therefore important to make it easy to formulate and refine queries, and to help the users understand the results. This is the main motivation for using phrasal captions and phrase matching. In outline, the processing in INVIP proceeds as follows. When images are registered with the system, their captions are analyzed into a meaning representation. The terms from the captions are stored in an index database, pointing to records containing the image identifier and the analyzed caption. In retrieval, the terms are extracted from the query and used to find candidate captions using conventional IR techniques such as vector-cosine matching and this phase is called simple matching. The query is analyzed to a meaning representation in the same way as the captions, and the representations of the query and candidate captions are compared using natural language matching techniques. The result of the comparison is a score, which is combined with the score from simple matching. Contexts may also be extracted at this stage, and the resulting images with their scores, captions and contexts are presented to the user.

Steganography can be used to hide phrases or captions in images stored in databases so as to make the search secure and fast. Steganography is more or less the same as the message encryption and message decryption in cryptography. It is this reason that steganography is often used together with cryptography. In general steganography system can either be secret or public. In public–key steganoraphic system different keys are used for message concealing" embedding" and message extraction[3],[4]. Natural language steganography techniques, which aims to hide secret information in text documents or photographs, by manipulating the semantic and/or syntactic structure of sentences. The objective of this paper is to propose a natural language steganography technique for hiding secret text

DOI: 10.37200/IJPR/V24I5/PR2020129
Received: 07 Mar 2020 | Revised: 26 Mar 2020 | Accepted: 04 Apr 2020

using LSB's. The proposal technique is different from all the previous methods in that cover carrier text is generated and translated into other language at final step. The implementation is made by four main steps and these are: the first step is random number generation (a large prime number is used to generate the secret key). The second step is steganography method implementation (The bits of secret message is embedded in the LSB's of the result secret key). The third step is text generation (the results stego carrier bytes is used to generate a text). The last step text translation (the result text is translated from English into any language).

3.2 Random Number Generator

Input: (p) prime number, k: bits number of secret text information, period length $l = n$,a,b

Output: (0 1 1 , ,..., $k - r\ r\ r$)

Process:

Step1: Initialization: input r,a,b,n 0 and k , $j \leftarrow 1$

Step2: Compute ($r_j \leftarrow arj_{-1} + b$)

Step3: $r_{i=}\ r_i$ (mod *number of the used lanugage*),and print $j\ r$

Step 4: increase j: $j \leftarrow j + 1$.

If $j \geq k$, then go to step 5, else go to step 2

Step 5: End

3.3 Proposal Hiding Algorithm

Input: Secret text information

Output: Stego-cover carrier text written in any language.

Process:

Step1: Convert the input secret text information into bits (*i mbit*).

Step2: Use a random number generator which initialized by a prime number (p) to generate (k) numbers (r_0 , $r_1 \ldots r_{k-1}$)

Step3: *i n* is the number result from Hiding (*i mbit*) into the LSB's of (*i r*)

Step4: *i w* is the letter result of Char (*i n*)

Step5: *i gfw* is the Generate forecast word beginning with letter *i w* .

Step6: Repeat step (1-3) until the end of (*i mbit*).

Step7: Use FOG, and *i gfw* to generate the bilingual text in English / French text using RECCE.

Step8: Translate the results form step 4, into any language "which represents the stego-cover carrier text".

Step9: End

IV. PHRASE MATCHING

Phrase matching is done by analyzing the query and the caption into dependency structures, in which the words are connected by labeled links indicating the relationship between them. One word (or occasionally more) will not be a modifier of any other words. It is designated the head, is the word which says, in most general terms, what the caption is about. The head of the query is compared against words in the caption, starting from its own head and progressing to modifiers if no match is found. If there is a match, the modifiers of the query head are compared

DOI: 10.37200/IJPR/V24I5/PR2020129
Received: 07 Mar 2020 | Revised: 26 Mar 2020 | Accepted: 04 Apr 2020

against modifiers of the corresponding term in the caption. Finally, allow matching of elements in the dependency structure against fixed expressions, to detect special cases such as negation. Figure-1 shows the dependency structures for two phrases with similar meanings. Dependencies are shown as pointing from a modifier to the term it modifies.

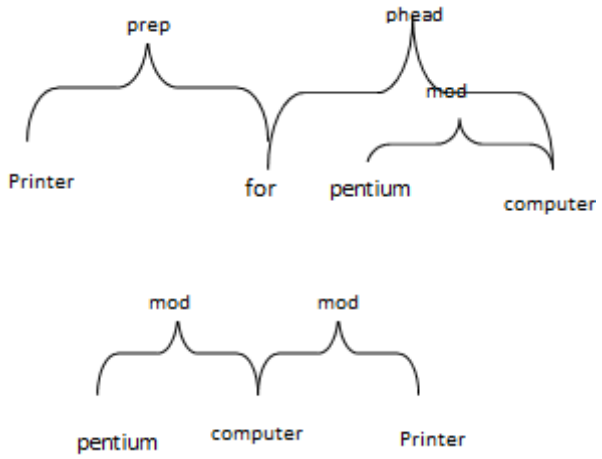

Fig. 1: Example Dependency Structures

For convenience, it represent dependency structures using a notation of indexed variables, in which the name of the variable stands for the name of the dependency, and the variable is indexed on the modified word. An un indexed variable is used for the head. The examples can then be written as color document copier

head = printer

mod[printer] = pentium

mod[pentium] = computer

printer for pentiumcomputer

head = printer

prep[printer] = for

phead[for] =computer

mod[computer] = pentium

Thus, for example, mod[printer] = Pentium indicates

thatprinter stands in the mod relation to pentium, i.e themod(ifier) of printer is pentium.

A system of rules specifies what relationships can be treated as equivalent. A small set of example rules appears in figure 2. The left and right hand sides of a comparison express paths through the dependency structure. The idea is that if we have already found a query word which matches a word from the caption, we then follow the specified paths from these words, and compare the words lying at the end of the paths.

4.1 The Scoring Scheme

The scoring scheme is a critical part of phrase matching, as it will allow us to distinguish exact and near exact matches from partial and weak ones. The general approach is to assign each word of the query phrase two numeric

DOI: 10.37200/IJPR/V24I5/PR2020129
Received: 07 Mar 2020 | Revised: 26 Mar 2020 | Accepted: 04 Apr 2020

values, called the score and the weight. The score of a query word is a measure of how well it matched considered in isolation from the rest of the caption, while the weight indicates the importance of the rule application. Writing the scores as si and the weights as wi, the overall score of the match is Psi wi/Psi, modified by the up-scores as described below. At the start of matching, the weight is 1.0. As we follow through continuations, it is the product of the factors of the rules leading to this point. If the rule above were in the start group, the weight of words matched in mod rule would be 0.7, and if mod rule contained

redbike
head = bike
mod[bike] = red
bike which is red
head = bike
rel[bike] = which
cop[which] = is
vhead[is] = red
bike which is not red
head = bike
rel[bike] = which
cop[which] = is
vhead[is] = head
amod[w] = not

Figure 2: Dependency Structures for the Matching Example

To show the rules in operation, suppose the query yellow car is tested against yellow car, car which is yellow and car which is not yellow. The dependency structures, written as variables, are shown in figure 3, and a trace through the matching process appears in figure 4. In particular, note how the rule'not' = amod[] 0.0 => Done 0.0;causes the previous score assignment for yellow to be replaced by 0 when comparing against car which is not yellow. The scores in this rule set are chosen on the basis of examining a variety of examples, some of which might be expected to provide a close match, some a partial match, and some a weak match. No experiments on learning the scores from data have been carried out.

V. EVALUATION

Evaluation of image caption retrieval is limited by the lack of suitable large test collections. Therefore we created our own captions for a set of digital photographs. The captions were prepared according to a set of guidelines, so that they emphasized the objects in the image rather than layout or composition. A query set was constructed by taking pictures from another source, and devising phrases which should elicit a related image. An initial set of results was obtained by pooling several keyword-based retrieval runs, discarding queries which produced no results. The top results from phrase matching with each query were then judged for relevance by two human assessors, acting separately. Neither assessor was responsible for writing the captions; one of them devised the queries. An example of the output for a query, showing some sample captions appears in figure-3. Phrase matching produces a good improvement over simple matching. 45 of the 55 queries in the best phrase matching run gave a precision of 100% at 10% recall. For example, the query tooth paste fail to match tooth brush because a brush is not normally a paste. The precision at 5 documents shows less of an improvement as a result of the small numbers of relevant captions.

DOI: 10.37200/IJPR/V24I5/PR2020129
Received: 07 Mar 2020 | Revised: 26 Mar 2020 | Accepted: 04 Apr 2020

Red bike +red bike

Query word	Rule group	Comparison	Score	Weight
Bike	Head_rule	Head=head	1.0	1.0
Red	Mod_rule	Mod[] = mod[]	1.0	0.7

Overall match score =(1.0*1.0+1.0*0.7)/(1.0+0.7)=1.0

Red bike + bike which is red

Query word	Rule group	Comparison	Score	Weight
Bike	Head_rule	Head=head	1.0	1.0
Red	Mod_rule	Mod[] = mod[]	1.0	0.7

Overall match score =(1.0*1.0+1.0*0.7)/(1.0+0.7)=1.0

Red bike + bike which is not red

Query word	Rule group	Comparison	Score	Weight
Bike	Head_rule	Head=head	1.0	1.0
Red	Mod_rule	Mod[] =vheadcop.rel[ ]	1.0	0.7
Not	Mod_rule	Not = amod[]	0.0	0.0
Red	Mod_rule	Mod=vheadcop.rel[]	0.0	0.7

Overall match score =(1.0*1.0+0.0*0.7)/(1.0+0.7)=0.50

Fig. 3: Matching in Action-(3)Cases.

Table 1: Evaluation Results

Run	Precision at 10% recall	Precision at 5 documents	R-Precision
Simple matching(I)	85%	45%	60%
Phrase matching(I)	92%	46%	66 %
Simple matching(II)	87%	49%	63%
Phrase matching(II)	96%	53%	72%

VI. CONTEXT EXTRACTION

Context extraction is a means of obtaining additional information about phrases which matched, by using the unmatched parts of the caption which are close in the dependency structure to parts which did match. For example, if the query was camera lens, and the captions included long camera lens and camera lens on a table, then the contexts would be long and on a table. Context extraction becomes valuable when there are many retrieval results. Captions with similar contexts can be grouped together, for example as shown the bottom half of in figure-5.

The algorithm for extracting the context is quite straight forward. Perhaps the most important point about context extraction is not the algorithm or exactly what the results look like, but the use of NLP to provide extract

DOI: 10.37200/IJPR/V24I5/PR2020129
Received: 07 Mar 2020 | Revised: 26 Mar 2020 | Accepted: 04 Apr 2020

information.

```
let P be the set of path rules (input)
let T be the set of current words, initialised to all matched words (input)
let U be the set of available words, initialised to all unmatched words (input)
let S be the set of contexts, intially empty (output)

while T is not empty
{
    select a word t from T

    for each word u in U
    {
       if there is a context rule <rt,rv,rp,ru,rC> in P such that
           has_pos(t,rt)
           AND in_var(t,rv)
           AND has_pos(u,ru)
           AND on_path(t,u,rp)
       then
           find the smallest phrase C such that valid_phrase(C,rC,u)
           if there is such a C then
               add the context <t,C> to S
               remove u from U
    }

    remove t from T
}

where
  has_pos(t,rt)   if t has part of speech rt
  in_var(t,rv)    if t is stored in variable rv
  on_path(t,u,rp) if the path rp connects t and u
```

Fig.4: The Context Extraction Algorithm

VII. CONCLUSION

The approach is the matching process looks for pairs of words which are syntactically related in the query tree, and which both appear in the tree for the key NLP(caption). The nearest parent nodes for the pairs of words are then checked for compatibility. The main way in which this differs from our algorithm is that the selection of nodes to try is adhoc ,rather than being guided directly by the modification structure. The use of rules with a reduced score (such as head = mod[] above) and mopping up rules is also more explicit and modular than the use of residuals. Furthermore, the scoring process in our phrase matching takes the depth through the structure (and hence the significance of the terms) into account better, and is arguably more perspicuous. Two challenges follow. The first is to adapt techniques of this sort to full text documents, in which there is a much richer linguistic structure, and where different parts of the text may have different information content (a title compared to a sentence in parentheses, for example). Secondly, there is a need to use evaluation measures which place more emphasis on interactive retrieval and user reaction. The assumption in much IR is that the results are simply judged by their relevance to the user's information needs, essentially as a binary decision. With an extension such as context extraction, where the retrieval

DOI: 10.37200/IJPR/V24I5/PR2020129
Received: 07 Mar 2020 | Revised: 26 Mar 2020 | Accepted: 04 Apr 2020

results contain extra information over the original data, it need an evaluation technique which is able to take into account the benefit obtained from the results by the information user. Natural language processing and steganography increases the efficiency and security of search and retrieval of data from image databases by using INVIP system, Phrase matching and context extraction method.

REFERENCES

[1] P. Selvi Rajendran,"Virtual Bulletin Board using Man-Machine Interface (MMI) for Authorized Users", *Indian Journal of Science and Technology,* 12(34), DOI:10.17485/ijst/2019/v12i34/112683, September 2019.Web of Science.

[2] T.P.Anithaashri, G. Ravichandran, R.Baskaran, Software Defined Network Security enhancement using Game Theory, Elsivere COMNET, vol157, pp:112-121, 2019

[3] P.Selvi Rajendran,"Virtual Information Kiosk Using Augmented Reality for Easy Shopping", International Journal of Pure and Applied Mathematics (IJPAM). special issue.Volume 118 No. 20 2018, 985-994,Scopus

[4] T.P. Anithaashri, G. Ravichandran , et.al. Secure Data Access Through Electronic Devices Using Artificial Intelligence, ICCES, 2018.

[5] T.P. Anithaashri, R. Baskaran, Enhancing Multi-user Network using sagacity dismissal of conquered movements, *International Journal of American Scientific Publishers* pp:69-78, 2016

[6] Deng, X, Deng, Y & Chan, FT. 'An improved operator of combination with adapted conflict', *Ann. Oper. Res.*, vol. 223, no. 1, pp. 451–459, 2014,

[7] Deng, X, Wang, Z, Liu, Q, Deng, Y & Mahadevan, S 'A belief based evolutionarily stable strategy', *J. Theor. Biol.*, vol. 361, pp. 81–86, 2014.

[8] TPAnithaashri and R Baskaran. Reign Monitor service for web enabled distributed system in the *International journal on Computation of Power, Energy, Information and Communication,* Vol-12 April 2013

[9] TP. Anithaashri and R Baskaran, Enhancing the Network Security using Lexicographic Game- Second International Conference, Advances in Computer Science and Information Technology-Bangalore, India, Part-III, Jan2-4,2012,.

[10] Florea, MC, Jousselme, AL, Bossé, E & Grenier, D. 'Robust combination rules for evidence theory', *Inf. Fusion*, vol. 10, no. 2, pp. 183–197, 2009,

[11] D. Elworthy. A finite set parser with dependency structure output in the 6th international workshop at Italy 2009

[12] Daugman, J. 'How iris recognition works', *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 14, no. 1, pp. 21-30, 2004,

[13] Lefevre, E, Colot, O & Vannoorenberghe, P 'Belief function combination and conflict management', Inf. Fusion, vol. 3, no. 2, pp. 149–162, 2002,

[14] Hespanha, JP & Bohacek, S. 'Preliminary results in routing games', *in Proceedings, American Control Conference*, vol. 3, pp. 1904-1909, 2001,

[15] Z. Shan, C. Lin, F. Ren, and Y. Wei. Modeling and performance analysis of a multi-server multi queue system on the grid. In Proceedings of the 9th International Workshop on Future Trends of Distributed Computing Systems, pages 337-343, 2000.

[16] S. Flank. A layered approach to NLP – based information retrieval in proceedings of 17th COLING at Montreal,1998

[17] Pollock and A. Hockley. What's wrong with internet searching? D-Lib magazine,1997

[18] A.F. Smeaton. Information retrieval: Still butting heads with NLP in Springer, 1997.

[19] A.F. Smeaton and I. Quigley. Experiments on using semantic distances between words in image caption retrieval in , proceedings of 19th SIGIR, pages 174-180 , 1996

[20] T. Strazalkowski. Robust text processing in information retrieval in the proceedings at the 4th conference at Stutgart, 1994.

[21] J.L. Fagan. Experiments in automatic phrase indexing for document retrieval. in 1987 at Cornell University

[22] C. Sindhu and G. Vadivu. Effects of Adjective Verb Adverb on Sentiment Classification Using Support Vector Machine for Green Communication. *Journal of green engineering,* volume.10,issue.1, pp.91–102, 2020

[23] M. Gayathri, C. Malathy, Siddharth Singh.MARK42: The Secured Personal Assistant using Biometric Traits Integrated with Green IOT, *Journal of green engineering,* volume.10, issue.1, pp. 255–267,2020.

DOI: 10.37200/IJPR/V24I5/PR2020129
Received: 07 Mar 2020 | Revised: 26 Mar 2020 | Accepted: 04 Apr 2020