International Journal of Information Technology and Business Management

29th January 2016. Vol.45 No.1

ISSN 2304-0777 www.jitbm.com

CONCEPT MAP CONSTRUCTION FROM E-COMMERCE WEB PAGES

Graziella M. Caputo[1], Nelson F. F. Ebecken[2]

[1,2]Federal University of Rio de Janeiro, Rio de Janeiro, Brazil

Email: [2]nelson@ntt.ufrj.br

ABSTRACT

Given the broad range of products and services offered by the companies, the choice of a specific item becomes a very arduous task. To decide on which item best meets the needs, it would be necessary to first know the background all the items available in the market and the technical details that surround them. This task, however, may be complex because the large number of choices and the limited knowledge of the customer. Therefore, this study aims to use information extraction mechanisms to better organize the details of products and services, creating simplified concept maps. Concept maps are mainly used in the learning process and therefore can offer tips that streamline mental absorption of knowledge. Much has been studied about automatic and semi-automatic construction of maps. In the present study, we applied text mining and natural language processing techniques to extract concepts and their relationships, to generate maps and maps comparisons.

Keywords: *Information Extraction, Conceptual Map, Natural Language Processing Knowledge management, Knowledge creation, Knowledge acquisition, Knowledge capture, Knowledge sharing.*

ISSN 2304-0777 www.jitbm.com

INTRODUCTION

Conceptual maps were developed to help students to better understand the main concepts involved in a specific subject. At the present work, the conceptual maps are used to help customers to better understand the main concepts involved in specific products, services or companies. As for students, costumers are learning about the new items they want to acquire. The amount of concepts that involves a specific subject depends on the complexity of each detail. Moreover, the velocity that a person will understand these concepts depends on a previous knowledge acquired about the subject. If the subject is new, it is timing consuming to take a decision to choose from a variety of offers.

On the other hand, if the user can easily find information about a specific product of a specific model, than, he can understand the implicit details. For example, how many colors exist for a specific product and the colors available for each model. From a business point of view, mechanisms to compare business information are very useful to provide competitive intelligence resources.

There are many strategies that a company can use concept maps to obtain valuable information, for instance, organize the web site concurrency to understand their offer or analyze their own website to understand how easy information can be found.

Nevertheless, there is a huge amount of documents that can be found in the internet or in the manufactories tutorials that tries to explain these concepts. However, each document will explain the same subject using different words. All these problems increase the time consumed to understand each subject. These issues can be solved using mechanisms to rapidly acquire knowledge and the concept maps can be used for this purpose.

To better extract the concepts in a big documents set it is necessary to use information extraction systems that can facilitate this process. It is necessary to use natural language process to disambiguate terms and create clusters of keywords that express the same idea.

The concept maps are widely used in learning techniques. It was developed by Joseph D. Novak (2010) to represent the students' knowledge in a specific domain. Nowadays, they are also used to help beginners in a subject to understand an idea, e-learning techniques and business environment, e.g. annotation, summarization, brainstorming and knowledge creation.

The methodology presented in this paper aims to interpret the information available in documents that describe products and create a mechanism to accelerate the information interpretation. To achieve this organization level, the work executes five main tasks: data acquisition and preparation, term extraction, concept extraction, relationship extraction and map comparison.

CONCEPT MAPS

Concept maps are used to organize and represent some knowledge about a specific subject. This representation is done by a graph structure where the circles represent the concepts and the edges the relationship between them. There are two challenges to construct the conceptual maps: the extraction of the concepts that describe the subject and the extraction of the phrases that represent the relationship that connect the concepts.

The maps are commonly constructed manually by a specialist. Each specialist of the same domain will construct a different map because it depends of each person perception. This indicates that it doesn't exist a correct map to represent any subject.

Doing this it is necessary much time and effort because of the amount of concepts that may involve a specific subject. With intelligent techniques evolution for information extraction, there are many studies that dedicate special attention on automatic and semi-automatic techniques for concept map construction. The initial idea is that the phrases can be broken in small pieces. Some small parts represent the concepts and while other parts represent the linkage sentence for the concepts.

ISSN 2304-0777 www.jitbm.com

The concept identification process implies the use of many linguistics resources that contains words information. It can be used ontologies Navigli et al (2004), glossaries and dictionaries to extract synonyms, relation between the words, like hypernym, hyponym and others.

According to Hu and Liu (2006) there are four main methods categories to find synonyms or lexical similarities between words: dictionary use, WordNet Fellbaum (1998), thesaurus and mutual information analysis of term pairs using co-occurrence Turney (2001). An important issue about products characteristics extraction is the amount of attributes that are generated, and many of them are synonym. For instance, the terms "photo" and "image" are synonym in digital camera subject, so they should be associated in the map construction. In this way, each context has particularities that should be taken in consideration during the subject analysis.

In Cañas et al (2003) it is presented an algorithm that uses the WordNet to perform the word sense disambiguation using a map that provides the context. The algorithm is based on 6 steps: key concept selection, words related with synset, hypernym sequence creation, clustering, best cluster selection and finally, word sense disambiguation. For Portuguese language, it was developed by Dias-da-Silva (2010) the Brazilian WordNet (WordNet.Br) containing many association between words like hypernym, hyponym, and synonym.

Usually, the methodologies to extract the concept two rules must be stressed: 1) it should be used just documents that belong to the specific domain, aiming to avoid ambiguity and 2) it should extract both multi-terms and single terms. For instance, in the sentence "digital photo frame", six terms should be initially considered: "digital", "photo", "frame", "digital photo", "photo frame" and "digital photo frame". This approach was applied in the present methodology.

On the other hand, [9]Jiang et al (2005) proposes an approach that increase multi-terms importance using documents with contrasting content aiming to train the classifier with negative content. It is developed a system called CRCTOL (Concept-Relation-Concept Tuple-based Ontology Learning).t combines statistical and lexico-syntactic methods, Word Sense Disambiguation and rule-based algorithm to extract relations and a modified generalized association rule mining algorithm that prunes unimportant relations for ontology learning.

Some automatic construction approaches consider that sentences can be broken in small pieces to find concepts. They usually analyze each paragraph, sentence or document separately, depending on the necessity. For this, it can be used grammatical or syntactic information. Using a grammatical parser, a sentence can be analyzed according to the grammatical tree that classifies sub-sentences as name or verbs. The sentence breaking uses punctuation or conjunction to separate pieces of sentences that will be considered as concepts.

In Tseng et al (2007), it is proposed an approach that constructs the conceptual maps in two phases (TPCMC – Two Phase Concept Map Construction). It uses historical testing records of students. The first phase pre-processes the records and creates the association rules using a fuzzy approach. The second phase transforms the association rules in concept relationships to finally create the map.

In Dias et al (2008), a methodology is proposed to automatically construct a hierarchical structure of names based on a bottom-up classification method. The internal nodes of the resultant tree have the hypernym of the grouped names using patterns like "B is a kind of A".

In Bai et al (2008), fuzzy rules are applied for the construction and evolution of conceptual maps using the relevance between the concepts. This work aims to improve the fuzzification performed in the work proposed by sue in Sue et al (2004).

According to Villalon et al (2009), the concept extraction can be performed in two

ISSN 2304-0777 www.jitbm.com

phases: the identification of possible concepts and the selection of the most important. This step is called summarization. The document D has all the words and phrases that potentially could be part of the ACME (Automatic Concept Map form Essay), that encompass the concept C, the relationship R and the topology G. In the work, it is formalized through D<={Cd, Rd, Gd} where Cd corresponds to all concepts on D, Rd corresponds to all relationship on D and Gd corresponds to all generalization level occurred in the document. According to this formalization, the concept identification correspond to identify Cd on D and the summarization consists of filter the Cd from C.

In semi-automatic approach, it is necessary the specialist help for the conceptual map construction. The system recovers information or suggests concepts and relationships. The user can use this information to construct the conceptual map filling the lack and correcting the mistakes.

In Hagiwara (1995) it is proposed an algorithm based on neural network called SOCOMs (Self-organizing concept maps). The algorithm suggests a concepts arrangement in the space using K-NN. It can use the information comparing the sentence or documents similarity, or using the ranking information from the similarity between items.

The semi-supervised method presented in [17] acquires patterns for each predicate (concept or relation), for example, in the sentence "the mayor of X" implies that X is a city. The algorithm achieves better accuracy using and adding more information to the knowledge base. For instance, if the algorithm finds some relation between two categories, it is verified, before adding to the base, if both instances are previously added in the base as the categories. This approach has the objective to maintain the information coherence.

METHODOLOGY

Information extraction (IE) is the process that extracts relevant data from a set of non-structured documents. There are many IE techniques that can be applied in context identification from specific business subject.

The methodology applied in the current work performs the process in five steps: data acquisition and preparation, term extraction, concept extraction, relationship extraction and Concept map comparison. . The figure 1 illustrates all the executed steps.

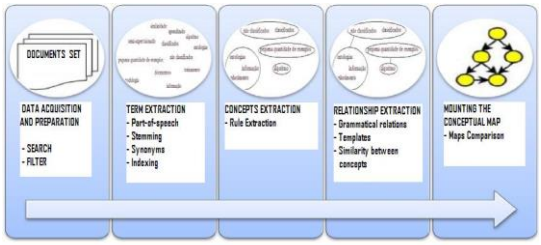

Figure 1: Proposed Methodology

As highlighted in [18] the business information in web pages are represented with the follow characteristics:

- Business information doesn't have a structural form and it doesn't have any logic sequence between the attributes.

- E-commerce stores usually don't provide all information.

- Some information are provided with many values, for example, phone numbers, like '02-555-1234, 1235. '.

- E-commerce stores provide information distributed over several web pages.

Data Acquisition and Preparation

The first phase searches on e-commerce web pages and filters the information considered relevant for the domain. It extracts the textual data from the internet based on some query. Many methodologies to execute a crawler can be applied depending on the knowledge that is aimed to be extracted. For example, if the study aims to understand a specific product (a model from a brand), it is better if it uses just documents related to that specific product. Otherwise, if the objective is to take a decision about witch product to buy from a range of available one in the market, it is

ISSN 2304-0777 www.jitbm.com

preferable to use the specification of all products that exist. If the objective is to contract a service, it would be better to compare the offers from all the companies.

If the internet page doesn't have many words, it will not be considered either. The searched pages were exclusively that ones that sell products or offer services related to the subject aiming to reduce redundancies. In this case, for services, just the web page of the company should be considered and discard other domains to avoid not useful data.

In this phase the information organization is maintained, that means, if it exist some table, bullets or titles, they was stored aiming to preserve the association between the terms. It is also performed data cleaning, to exclude HTML tags and term correction, where misspelled words could exist, generated from the specific language used by the web pages.

Term Extraction

In the term extraction phase aims to identify the single terms and multi-terms that represents the data. The multi-terms occur when a set of terms are frequently close to each other. Then, it is considered the frequency of each word and the co-occurrence between them. This phase also takes in consideration grammar structure.

The analysis phase starts observing each term individually, storing the frequency and the position of each of them in the document. For instance, consider the two terms term1 and term2, representing the concepts "photo" and "digital photo frame" respectively. It is most probable that the frequency freq_term1 of the first term is greater than the frequency freq_term2 of the second term. Otherwise, both frequencies will be the same, once term2 contains term2. In other words, freq_term1 ≥ freq_term2 if term1 is contained in term2. However, if term1 occurs with high frequency in the set of documents, its relevance will be lower.

To determine if a set of words is a term or not, it should considerate grammar structure and natural language processing techniques (NLP). In term extraction step, the tasks performed are: Part of Speech (POS), stemming, synonym (WordNet.br) and indexing.

The POS identifies if the word is a verb, noun or adjective. If a set of nouns co-occurs frequently, they are considered as a unique term. The same is done for the adjectives. Usually the verbs represent the relationship between the nouns and/or adjectives, so they are analyzed just in the relationship extraction phase.

The stemming extracts the stem of each word inside a term. The stem is stored for each word with the original POS classification. For instance, it is important to maintain the POS classification because some words can have classification ambiguity. For example, the words "provide" (VERB) and "provider" (NOUN) has the same stem: "provid".

The indexing methodology used in this paper intends to save the document structure to make possible the position analysis of each word and distances between them.

As an example, in the phrase "the value of the barrel of oil is more expensive" can be reconstructed as "The (SW) value (N) of (SW) the (SW) barrel (N) of (SW) oil (N) is (V) more (ADV) expensive (ADJ)". Where N is noun, V is verb, ADV is adverb, ADJ is adjective and SW is stopword. In this case, the names are being compared with each other even though they have stop-words between them, because the distance measure does not count SWs. In the example, both "value of the barrel" and "barrel of oil" terms were considered to possess an equal distance. They can be considered as terms because the words inside the sentence have the classification N (excluding SW). Finally, in the sentence "The value of a barrel of oil is more expensive," the term "value of a barrel of oil" will be considered as a term. Once more, the most important terms are maintained, otherwise, they are excluded.

Concept Extraction

In this phase the terms that represent the same concept are grouped in the same concept. To do so, a set of rules were created to reduce the number of terms with the same

ISSN 2304-0777 www.jitbm.com

meaning, which means redundancy.

All the terms considered with the same meaning were grouped in this phase and considered as a concept. At the end, the number of concepts is smaller than the number of terms. The resultant table stores the concepts with their position in the documents. The following rules were considered to group concepts:

° Numbers were replaced by %num%, and consequently the terms that differentiate Just by number were grouped together. E.g. "40 inches TV", "32 inches TV" were grouped in the term "%num% inches TV";

° Terms that differentiate just by stopwords

° Words in different sequence inside a term

° Repeated words inside a term

° Grammatical Analysis – frequent relation between adjective and noun.

Relationship Extraction

Once identified the concepts contained in the document collection, the next step aims to identify the concept relationships.

To extract the relationships, three approaches were applied: grammar relations, the document structure and concepts that groups terms with common differences.

The grammar relation between the concepts is the principal analysis to be made. The verbs can be considered as indicators for semantic relations. It can be said that the relation NounVerb-Noun can be mapped to the relation Concept-Relationship-Concept. If a concept is frequently related to another concept, given a small distance in a set of text, it can be said that they have some relationship. The presence of a specific verb within these concepts is a good indicative of the relation between them.

In this case, it is considered the frequency that the tuple Concept1-Verb-Concept2 had occurred. It was not taken in consideration in how many documents it happened but just the total amount of co-occurrence.

The second approach for relation extraction considers the formal structure of the web pages that facilitate the relationship association. In the product case, it is common the use of tables where one column presents the characteristic type and the second column presents the description. The Table 1 describes an example of this structure for a television presentation in a web site.

The third approach considers the uncommon terms inside a concept. For example, the previously grouped terms "Digital Photo Frame 8 inches" and "Digital Photo Frame 9 inches" (the "#num" is replaced by the original number) indicates that it may have two sizes for digital photo frames: 9 and 8 inches. This can be considered that there is a relation between digital photo frames and the number of inches. Note that it was previously considered as a unique concept because both are nouns and co-occurs frequently closed.

Table1: Table Specification

Characteristic	Description
TV Type	LCD Flat-
Screen Size Class	46"
Internet	No

Map Comparison

The map comparison has the objective to show the difference between the compared subjects. It means that if the objective is to compare two products, the functionalities of both will be shown in the map. Otherwise, if the objective is to compare two companies or services, the details will be compared there.

First of all, the map of each context is created linking the concepts using the relationship. Once you have the two concept maps , concepts of objects are marked with a flag being set, for example "A" for the concepts of an object, and "B" to the concepts of another object. Those concepts that occur with both flags are changed, for example, "C".

The concepts that belong to both subjects are grouped in the same node and it is preferable that each subject have a different color to facilitate the rapidly understanding of the

ISSN 2304-0777 www.jitbm.com

difference.

The main problem on this case is when the different web sites present the same concepts with different terminology. In this case a specialist should validate the map.

EXPERIMENT

This case study deals with data available on the Internet that relate to companies sites that provide telephone services in Brazil.

The main challenge is the wide variety of issues that each company offer and have different way to be presented to customers. These differences are marketing strategy, where the company name is presented as part of the name of the service. Moreover, the fact that they have different service offerings is a challenge to create a single map that represents the type of company.

The methodology was applied on two companies web sites witch focus are to sell mobile services. All domains that belongs to the original home page was crawled. In total 10.224 documents were downloaded from both companies.

The first phase, the data cleaning phase, filtered documents that had no meaningful content, that means, not enough information. Thus, 7.804 documents were maintained.

The extraction phase aimed to examine each term individually and them the relationship. The initial documents set had 39,759 words, including grammar variations , as plural and verb conjugations. The stopwords and numbers were not considered.

The POS analysis identified 21.418 NOUNS, 3.067 VERBS and 4679 ADJECTIVES. In some terms the POS classification were not identified. The stemming algorithm were applied to replace to the original stem, maintaining the POS information.

A Table is constructed to show the simple terms with frequency and amount of documents in which they occur. The co-occurrence of all terms were analyzed.

The extraction phase of concepts used synonym dictionaries for rules identification and application for identifying patterns in sentences. Words representing synonyms were grouped into a single term, such as “web” and “internet”.

Furthermore, comparison rules were applied for words inside a term, such as order, difference by just stopwords and numbers, word repetition within the terms and others. A Table lists some terms identified by these rules and that were grouped togheter.

For the relationships extraction, the data had to be divided into two databases, one for each company. The first analyzed company offers a wide range of services.

The last Table presents the concepts that are directly related to the companies. They have a variety of offers that are presented in different ways. The table shows the terms that were grouped into concepts that have the company name included.

Both companies have the same naming strategy services, including the company name in the subject matter of the offer, such as "Oi Internet" and "Claro Card" to indicate internet services and card plan, respectively. The companies also employ different names for their products, such as the company calls its broadband as "Velox". Thus, it was analyzed the main concepts related to the subject, in the relationship phase.

The concepts directly related to the main term were extracted and them the related to these terms, and so on to create the concept map.

For this analysis two concept maps were initially created to show the specification of each company. Because they use names to specify their products, the similarities were clearer in more deep levels. The concept map for both companies can now be easily built.

Once established the main concepts of both companies, and the relationships between them, both maps were compared and unified. So that the concepts that occur in both were highlighted with a color, and the concepts occurring just for one single company were highlighted with different colors.

Figure 2 shows the information integration

ISSN 2304-0777 www.jitbm.com

from both companies. In this map it is possible to see the different market strategies, such as the inclusion of a seminar sponsored by a company or product promotion for another.

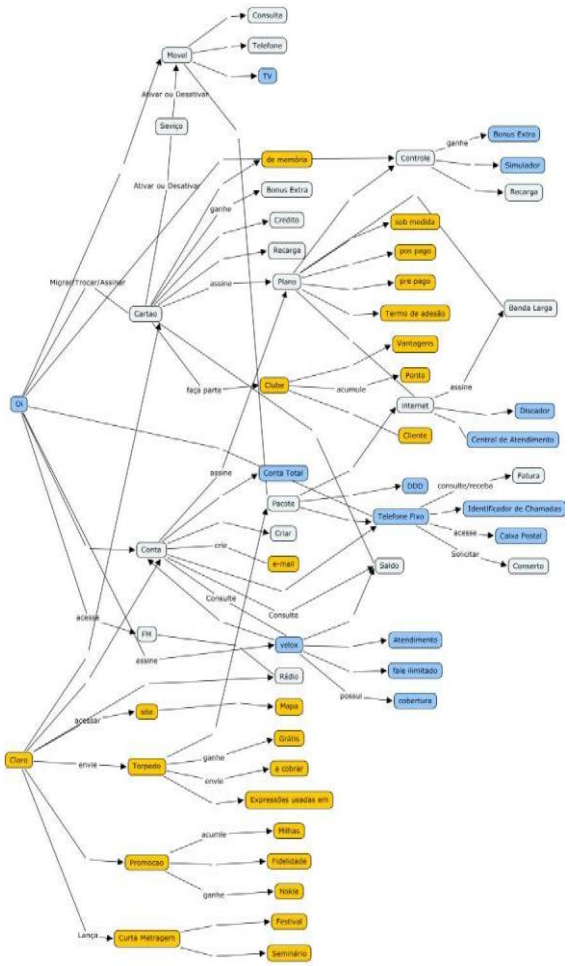

Figure 2: Conceptual Map comparing

CONCLUSIONS

This methodology creates concept maps in a semi-supervised way to specific domain documents. Using this methodology, it is possible to create maps for two specific companies and then, compare the differences between them. And it is also possible to create a unique conceptual map to describe the domain, using documents of all kinds of products.

The concept maps creation for business service providers resulted in three different maps, two to represent companies' strategies and one to compare them. The results could highlight the difference between the two without being initially generated a model for service companies.

It was executed four phases that collect the data, prepare and extract the information aiming to speed the learning process of specific subject. This methodology can also be expanded for many kinds of textual data that describe specific technical information.

ACKNOWLEDGMENT

We would like to thank the National Research Council, CNPq, and FAPERJ the Rio de Janeiro State Research Foundation for financial support.

REFERENCES

1. Novak, J.D. (2010). Learning, Creating, and Using Knowledge: Concept Mapas as Facilitative Tools in Schools and Corporations (2nd Ed.) NY: Routledge.

2. Navigli, R.; and P. Velardi (2004). Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites, Computational Linguistics, vol. 30, pp. 151--179.

3. Nakagawa, H. and Mori T. (2002). Simple but powerful automatic term extraction method. In COMPTERM 02, pages 29–35.

4. HU, M. and LIU, B. (2006). Opinion Extraction and Summarization on the Web. Proceedings of 21st National Conference on Artificial Intelligence (AAAI-2006, Nectar paper), July 16.20, Boston, Massachusetts, USA

5. Fellbaum, C. (ed.) (1998). WordNet: an electronic lexical database. Cambridge: The MIT Press.

ISSN 2304-0777 www.jitbm.com

6. Turney, P. (2001). Mining the Web for synonyms: PMI-IR versus LSA on TOEFL, ECML-01.

7. Cañas, A. J., Valerio, A., Lalinde-Pulido, J., Carvalho, M., & Arguedas, M. (2003). Using WordNet for Word Sense Disambiguation to Support Concept Map Construction. Paper presented at the Proceedings of SPIRE 2003: International Symposium on String. Processing and Information Retrieval, Manaus, Brasil.

8. Dias da Silva, B. C. Brazilian Portuguese WordNet: A Computational Linguistic Exercise of Encoding Bilingual Relational Lexicons (2010). International Journal of Computational Linguistics and Applications, New Delhi, v.1, n. 1-2, p.137 - 150.

9. Jiang Xing, Ah-wee Tan (2005). Mining Ontological Knowledge from Domain-Specific Text Document. Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM'05).

10. Witten, I. H.; Paynter, G. W.; Frank, E.; Gutwin, C.; Nevill-Manning, C. G. (1999). KEA: practical automatic keyphrase extraction, in Fourth ACM conference on Digital libraries.

11. Tseng, Shian-Shyong; Pei-Chi SUE, Jun-Ming SU, Jui-Feng Weng, Wen-Nung Tsai (2007). A new approach for constructing the concept map. Computers & Education Vol. 49, Issue 3, pp 691-707. Nov.

12. Dias, G., Raycho M. and Guillaume C. (2008) Mapping General-Specific Noun Relationships to WordNet Hypernym/Hyponym Relations. Springer-Verlag Berlin Heidelberg. pp. 198 – 212.

13. Bai, Shih-Ming; Chen, Shyi-Ming, (2008). Automatically constructing concept maps based on fuzzy rules for adapting learning systems. Expert Systems with Applications. Volume 35, Issues 1-2, July-August, pp 41-49

14. Sue, Pei-Chi, Jui-Feng WENG, Jun-Ming SU, and Shian-Shyong Tseng, (2004). A new approach for constructing the concept map. In Kinshuk, Chee-Kit Looi, Erkki Sutinen, Demetrios G. Sampson, Ignacio Aedo, Lorna Uden, and Esko K ahk onen, editors, ICALT. IEEE Computer Society.

15. Villalon J, CALVO RA (2009). Concept Extraction from student essays, towards Concept Map Mining. Proceedings of the 2009 Ninth IEEE International Conference on Advanced Learning Technologies - Volume 00: 221-225.

16. Hagiwara, M.. Self-organizing concept maps (1995). In IEEE International Conference on Systems, Man and Cybernetics. Intelligent Systems for the 21st Century, volume 1, pages 447{51, New York, NY, USA.

17. Carlson, A., J. Betteridge, B. Kisiel, B. Settles, E.R. Hruschka Jr. and T.M. Mitchell, (2010) Toward an Architecture for Never-Ending Language Learning. In Proceedings of the Conference on Artificial Intelligence (AAAI).

18. Sung, Nahk Hyun; Chang, Yong Sik, (2004). Business information extraction from semi-structured webpage. Expert Systems with Applications. Volume 26, Issue 4, May, pp 575-582