Concept Maps Construction Based on Exhaustive Rules and Vector Space Intersection

Josiel C. Reis,[†] Antonio S. C. Gaia[††] and Raimundo Viegas Jr.[†]

[†] Federal University of Pará, UFPA , Belém, PA, Brazil
[††] Federal Institute of Science and Technology, IFPA, Belém , PA, Brazil

Summary

This paper presents an alternative automatic method for the tuple-fashioned extraction of concepts and their binding elements aiming the construction of Concept Maps out of Brazilian Portuguese web texts, based on combined techniques of automatic generation of exhaustive syntactic rules, restricted-context part-of-speech tagging and vector space intersection. The results showed the proposed method *automatically* produces high-granularity propositions consequently elevating the semantic quality of the extracted binding elements. The method is able to discard the use of stemming techniques – in the specific scenario – before attributing weights to candidate terms.

Key words:

Concept Map, Text Mining, Natural Language Processing, Information Extraction.

1. Introduction

The overwhelming proliferation of information on the Internet has raised in academic, government and entrepreneurial areas the need for constant research on sensitive information extraction, text summarization and adequate information mining techniques, according to purposive areas of interest [2]. Also, several techniques and methods have been exploited in the last-decade academic world in order to improve knowledge representation techniques [10][11][12]. In this context, Concept Maps - originally conceived as an evaluation tool for assessment of meaningful learning patterns [12] – have also been revealed as an alternative suitable graphical tool for representing summarized knowledge of any chosen domain presented on Web texts [2]. Nevertheless, most of the experiments related to automatic or semi-automatic generation of concept maps out of unstructured texts, while dealing with Natural Language Processing issues and the problem of election of terms to be used as concepts or linking elements, make use of limited and specific list of rules which tend to reduce the binding elements to a very-few-terms compound string or even to a single stemmed term [5][13][14]. On the other hand, even in those (semi) automatic approaches where granularity levels of extracted information are higher, the extraction, the election and even the very proper linking of concepts may require considerable interference and effort of the user [1][2]. That scenario, though suitable for educational-purpose approaches, usually results in a limited and inadequate approach when dealing with the automatic extraction of critical quality information out of news or automatic good-quality text summarization or knowledge representation, due to loss of quality in much of the original semantic information. This paper provides refining improvements on the work of [15] for part-of-speech tagging on Brazilian Portuguese corpora, as well as presents a simplified automatic method on the process of obtaining more meaningful sentences for the construction of concept maps, and combines automatic generation of exhaustive lexical rules from the analysis of the very corpus to the selection of final tuples by resulting (dis)similarities from vector intersection process. In Section 2 we present the theoretical scope considered in the experiment. In Section 3 Related Works are discussed. In Section 4, the Methodologies and the Experiment are described. Section 5 concludes the paper and Section 6 discusses related future works.

2. Theoretical Scope

2.1 Text Mining and Information Extraction

The text mining techniques are assumed to be applied to any set of documents in diverse knowledge domains. The basic issues on text mining are related to NLP. In [3], a list of top ten issues on text mining is presented, ranging from stop words to stemming to tokenization and word tagging. Linguistic methods have also been heavily used and are highly related to the structural text characteristics and are mainly based on linguistic patterns identification and on detection of terms, using syntactic rules [9]. On the other hand, statistical techniques are intended to produce information and allow term frequency analysis and co-occurrences between terms in a corpus or corpora [9]. The most common text mining approaches deal with probabilistic or rule-based solutions or even hybrid solutions in order to assess unstructured

Manuscript received July 5, 2014
Manuscript revised July 20, 2014

information. [3] refers that text mining differs from information extraction (IE) because text mining regards the search through unsuspected unstructured information. In IE, information is predefined and matches the user interests. Nevertheless, IE techniques have become part of text mining tasks on behalf of knowledge extraction. As for parsing, in [15] resources are presented for developing a part-of-speech tagger with a 99% accuracy level regarding Brazilian Portuguese texts. And [4] presented the first reported attempt on using improvements on [15]´s work in order to use the resulting tagged corpus for information extraction aiming a concept map construction. Nevertheless, [4] focused only on a very restrict domain corpus and the experiment reported failure in extracting complete information and cardinality representation from texts.

2.2 Concept Maps

Concept Maps (CMaps) are basically graphical tools for knowledge representation. They were proposed by J. D. Novak [12][13] and their theoretical basis relies on David Ausubel's Assimilation Theory [6]. Concept maps are usually structured in a hierarchical distribution [9]. The most inclusive concepts occupy the top positions on the map. The less inclusive ones are intended to represent a more detailed category of concepts which occupy a descending position on the graph. Propositions are the basic constituents of concept maps and are formed by two concepts linked by a word or a set of words, in order to express a meaningful statement [12].

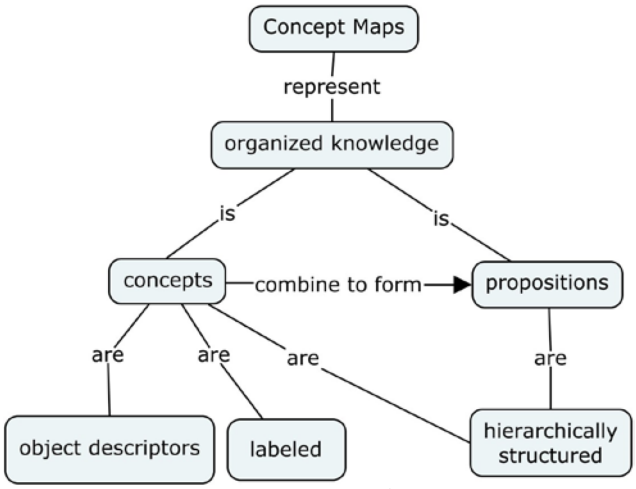

Fig. 1 A Concept Map adapted from Novak [12]

Main motivations for adoption of Concept Map approach are rooted on educational ground. So, from its conception the concept maps have been widely applied for educational purposes [11], mainly for the assessment of meaningful learning patterns [12]. Some other researches have been inspired by the need to acquire knowledge domain [5] and for exploration of digital documents sharing. [2]. In this paper Concept Maps are seen as a vital prominent means of knowledge representation regarding the results of text mining techniques applied on Brazilian Portuguese texts extracted from Web news. Nevertheless, the solution is expected to be effective on any domains of formal Portuguese speech. The texts to be submitted to the method are assumed to be well-written texts and are supposed to meet Brazilian Portuguese syntax rules' standards.

3. Related Works

In [2] an experiment is conducted on extracting information from web documents in order to construct a preliminary conceptual map for human refinement. The experiment is based on an 8-steps procedure which comprises a range from segmentation and parsing to normalization through stemming of terms to ranking of individual words and composed concepts, to concept extraction and selection. The experiment produces a sub-tree for each noun phrase which elevates the algorithm complexity. Though it uses simple term frequency for the term weighting step, on the other hand it also consults external thesaurus for synonyms. The method produces isolated concepts not used in the construction of the concept map. The experiment reports not being able to use some prepositional forms for linking phrases. In [13] strong participation of the user is required to obtain quality in the CMap The user is prompted to provide up to 5 subjective weights to classify the text. The tool aims detecting relevant terms to be used in CMap construction, to represent courses. The paper presents no resulting CMap and does not list the terms retrieved by the tool. [9] Describes a set of characterizations for CMap constructions but the paper presents no formal tool solution. The evaluation is conducted on directed-domain texts only. In [14] an experiment is presented where the creation of courses is represented by CMaps. The solution is based on 5-steps processing and goes through parsing, tokenization, pattern recognition and semantic analysis. The tool is applied on specific directed domain limited to the construction of CMaps related to some desired courses. In [16] the extracted terms are aggregated into a "proximity array" in an attempt to translate student essays into CMaps. However, only "concrete concept" terms are extracted producing a low granularity CMap. [4] presented the first reported attempt on using improvements on [15]´s work in order to use the resulting tagged corpus for information extraction aiming a concept map construction. Nevertheless, [4] focused only on a restricted domain corpus. Moreover, the experiment reported failure in extracting complete information and cardinality representation from texts. In our approach, the solution has

shown to be able to successfully extract high-granularity information from a wide range of domains (meteorology, sports, economics, biology and tourism are some of the explored domains). The user is not prompted to input any subjective weighting factors. The final tuples are automatically chosen and presented to the user, although the user may opt for downsizing the final tuples set presented, for convenience.

4. Methodologies and the Experiment

4.1 Tokenization and Parsing

Part-of-speech tagging usually requires prior preprocessing steps where the unstructured text must be conformed to the text format accepted by the parser tool to be used. In our approach, part of application includes Java programs which reflect post-processing improvements on the work of [15]. In [15] a two-step preprocessing is performed and the steps are described. The first step (preprocessing1) deals with strings replacement where compounded words (adjectives, prepositions and nouns) are disjointed and split in two or more terms to be properly parsed along the next step, and spaces are replaced by tab stops. The second step (preprocessing2) uses tab stops as reference to apply tokenization on the previous step results in a way that only a single token per line (word, punctuation signs and graphic symbols) is allowed, and gives to the file the format accepted by Tree Tagger parser. The next step is submitting the resulting file to Tree Tagger parser – so that each word in the text will be tagged by a lexical ID (noun, verb, adjective and so on) using parameters set to Portuguese language (portuguese.par). The "portuguese.par" file is a trained corpus extracted from CETENFolha[1]. By the use of the parameters file, the submitted text is then labeled in one-tagged-word-per-line format. A comparison step is run by comparing the obtained word-label pairs to the previously labeled pairs in the trained corpus, and by identification of the matching pairs (hits) and non-matching pairs (errors). The accuracy is then calculated as the sum of all hits/errors divided by the total number of word-label pairs submitted to labeling. In order to retrieve VERB-NOUN-VERB (VNV) triples [15] uses an editable Java class to be compiled and run by command line in order to save the obtained triples into a text file. Our contribution to this first stage is a refined post-processing mechanism which improved the results of [15]'s original work by extending the handling of exceptions to those not included in the original set of tagging rules. Composed words and expressions in Portuguese have received special attention (e.g "Copa_do_Mundo", "em_função_de") . Under the same scenario[2] and the same corpus to be parsed, (140 mutually-distinct words in a 263-words text) the solution in [15], as it is, produced a single triple under the VERB-NOUN-VERB (VNV) lexical rule, and no triples under NOUN-VERB-NOUN (NVN) lexical rule. By comparison, the modified text-mining process – which was improved by the new set of exhaustive rules applied to our approach – produced 23 final tuples in the NOUN-VERB-NOUN (NVN) format, and 2 tuples in the VERB-NOUN-VERB (VNV) format. It should be noted that a whole set of tuples was retrieved (282 raw tuples) which is not only based on or limited to the above-mentioned formats. After the intersection technique was applied, the final set was reduced to only 87 tuples. The solution is integrated in a web-based tool to be referred as *TagMiner*. The tool embeds two java classes to be locally run. Nevertheless, a web-based module has been developed to comprise both a user graphical interface and a tuple-extraction module which works on the java classes resulting files.

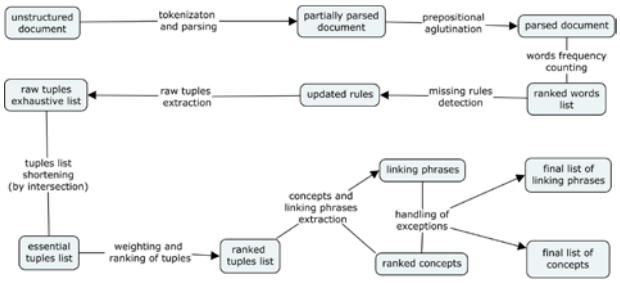

Fig. 2 The Cmap tuple-extraction cicle

4.2 Exhaustive Lexical Rules

In text mining, the quality of the retrieved set of tuples is strongly dependent on the lexical set of rules applied to the process. Starting from a basic set of simple syntactic rules (NVN, VNV) we have expanded the concept of noun (N) to include other syntagmatic components. Proper names and – in some cases – past-participle verbal forms are part of the exceptions included in that category. As for the binding terms, not only verbs or prepositions were considered, but also the syntactic constructions that give continuity to the semantic meaning of coordinate clauses - either at the beginning or in the middle of sentences (eg: the Portuguese adverbial locution *"no entanto"* which may signify *"yet"*). To the extent that new domains came to be explored, there arose the need of adding new syntactic rules to the set of rules. So, if a given syntactic construction is found on the text but is not yet in the set of rules, an algorithm is responsible for adding the novel rule to a temporary set of rules. And the process is repeated continuously until every new syntactic construction found is included in the set of rules. Rules are composed of up to four predicates. A *k*-predicate-based rule should be formally expressed by :

$$R=\{p_i,\ p_j, p_q, \ldots, p_t\} \quad (1)$$

Where R stands for the set of *k*-predicate rules, and each p_k identifies a lexical ID tag (eg., N, V, ADJ) which can be combined into an arrangement of up to four predicates to form a valid rule. A typical set of rules can also be expressed in the form of Boolean statements:

$$R= (p_{i="N"} \text{ or } p_{i="NPROP"}) \text{ and } (p_{j="V"} \text{ or } p_{j=}\ldots) \text{ and } (p_{q="ADJ"} \text{ or } p_{q="PCP"}) \text{ and } \ldots \quad (2)$$

Where R once again represents the set of combined *k*-predicates rule. If a "bad" rule is added to the set of rules based on a not "well-written" document, chances are very low that it will be found again in another document, so it will do no great harm to the extraction of its respective set of tuples. The more complete the set of exhaustive rules, the more complete set of high-granularity tuples shall be extracted.

4.3 Vector Intersection

Let there be two line vectors U and V which are defined as subspaces to R^x, where $x = \{1,2,3,4\}$. Let U and V be defined as a set of indexed terms (words). So:

$$U=\{\text{"credit", "portability", "program"}\} \quad (3)$$
$$V=\{\text{"starts", "Monday"}\} \quad (4)$$

Let the index or "weight" to each element *u* in U be the very position *j* of each element *u* in line *i*=1, on the line vector u_{ij}. So, for *j*=1, the value of u_{11} = "credit"; u_{12}="portability" and so on. In analogous way v_{11}="starts" and so on. If we consider W, as being the resulting set

$$W = U \cap V = \{\emptyset\} \quad (5)$$

then U and V are called orthogonal vectors and W is a subspace vector, called a null vector as the result to the intersection between U and V. If we consider U and V as sets of tuples to be compared, their eligibility for final tuples will be inversely proportional to their (dis)similarity. In short, only sets of tuples whose intersection to another tuple set results in a null vector W={∅}, are eligible to be added to the final tuples set for the construction of the corresponding Cmap. That way the redundant very-similar tuples are discarded. The well-known mathematical approach presented by [7] could be used here as well , though that solution is more suitable for calculating similarities between a group of different documents with different term weighting factors. In this paper we have opted for a simplified term weighting method as in [2].

4.4 Term Weighting

Word frequency counting is only a step away from parsing (see Fig. 2). For the specific case, as explained in the previous paragraph, a simple term-weighting method is preferred, since we are trying to obtain a single Cmap from a single document and there is no need for considering information from other sources which would require the use of Salton´s [7] approach. In this case, the very simple frequency of each word is attributed as its individual weighting factor. As for the tuples, their individual weighting factor consists of the sum of each of their constituent word's frequency. As a matter of fact, as a strategy to obtain the higher possible granularity, any word with a frequency higher than 1 is ranked and consequently affects their tuples' ranking. No stemming step is applied neither, once we intend to maintain semantic integrity. Stemming reductions may cause loss of semantic integrity in this specific case, since the sentence "Mary – *be* – married to John" allows ambiguity if compared to "Mary – *was/is* – married to John". The next step consists of electing concepts and linking phrases from the final tuples. This step requires attributing to each word an index related to its preceding pair. So, whenever the algorithm elects a concept, its preceding and posterior linking phrases are also retrieved in order to guarantee semantic completion to the final proposition.

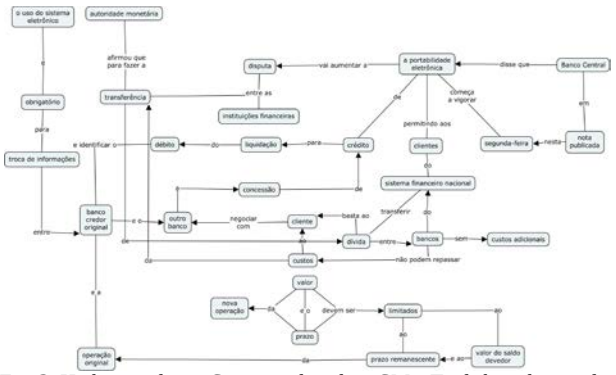

Fig. 3 High-granularity Cmap produced on CMapTools based on tuples extracted out of text mining on Web news

4.5 Handling of Exceptions

Finally, some exceptions must be properly addressed in order to obtain the clearest propositions without loss of significance. Due to the hierarchical characteristic of Cmaps, the order of a concept in a proposition might demand some adjustments. As an example, take the sentence *"É obrigatório o uso do sistema"* ("It is mandatory to use the system"). In this case, the nucleus of the noun phrase ("uso do sistema") has been moved to the end of the sentence. The algorithm must identify the inversion by checking the verb and noun positions. After adjustments, the new proposition should read "*O uso do sistema é obrigatório*" (whose translation to English allows "Using the system is mandatory". Although legibility was hampered by the size of the

picture, in (Fig. 3) we present a Cmap constructed by CmapTools[3] on final tuples extracted from Web news.

5. Conclusion

In this paper we presented a simplified method for extracting information from Web news, using text mining techniques, and introducing the self-produced exhaustive rules technique in order to obtain high-granularity quality Cmaps. Though we have used wen news as a source of information the presented method has been successfully applied to any well-written documents of multiple knowledge domains. We also presented the automatic extraction of final tuples by the use of vector intersection method which saves effort from the user. It is up to the user, the option of downsizing the final set of tuples according to convenience. High-granularity maps are intended to summarize information guaranteeing semantic integrity of original information as much as possible. Its direct benefits are related to elevating the information granularity levels, and, consequently, maintaining the semantic levels of integrity for the original information and high quality concept maps construction.

6. Future Works

As future works we intend to implement a new module to *TagMiner* in order to be possible for the user to visualize automatic independent Cmap generation following the presentation of final concepts and linking sentences. In addition we also noted that some concepts are set apart by semantic reasons, and some of them could be properly unified in a single related concept. A next step could be developing a semantic search which could be able to identify close-related concepts in order to obtain more concise and clear Cmap construction.

References

[1] Alberto J. Cañas et al (2004). Mining the Web to Suggest Concepts during Concept Map Construction. In A. J. Cañas & J. D. Novak & F. M. González (Eds.), Concept Maps: Theory, Methodology, Technology, Proceedings of the 1st International Conference on Concept Mapping. Pamplona, Spain: Universidad Pública de Navarra.Pamplona.

[2] Alejandro Valerio & David Leake, (2006) "Jump-Starting Concept Map Construction with Knowledge Extracted from Documents", Second International Conference on Concept Mapping, San José, Costa Rica Sept. 5-8.

[3] Anna Stravrianou, Periklis Andritsos, Nicolas Nicoloyannis (2007) "Overview and Semantic Issues of Text Mining". SIGMOD Record, September 2007 (Vol. 36, No. 3)

[4] Clay Palmeira, Rafael Chaves, Hamilton Cavalcante, Eloi Favero (2012), "A Requirements Elicitation and Analysis Aided by Text Mining". IJCSNS – International Journal of Computer Science and Network Security. Vol. 12, No 6 pp. 122-128.

[5] Cláudia Camerini Corrêa Pérez, Renata Vieira (2005) "Mapas Conceituais: geração e avaliação". In: TIL - Workshop de Tecnologias da Informação e Linguagem Humana, 2005, São Leopoldo. Anais do XXV Congresso da SBC. Porto Alegre : SBC, 2005. v. 1. p. 2158-2167.

[6] David Paul Ausubel, (1968), "Educational Psychology, A Cognitive View". New York: Holt, Rinehart and Winston, Inc.

[7] Gerard Salton & Christopher Buckley, (1988). "Term-weighting Approaches in Automatic Retrieval". Information Processing and Management. Vol 24, Number 5.

[8] Harry Kornilakis, Kyparisia A.Papanikolaou, Evangelia Gouli, Maria Grigoriadou (2004a) "Using Natural Language Generation to Support Interactive Concept Mapping". Department of Informatics & Telecommunications, University of Athens.

[9] Juliana H. Kowata, Davidson Cury, Maria Claudia Silva Bôeres (2009) "Caracterização das Abordagens para Construção (Semi) Automática de Mapas Conceituais". Universidade Federal do Espírito Santo (UFES). XX Simpósio Brasileiro de Informática na Educação (2009)

[10] Joseph D. Novak, D. Bob Gowin, Jane Butler Kahle (1984). "Learning How to Learn". New York: Cambridge University Press.

[11] Joseph. D. Novak.: Learning, creating and using knowledge: Concept maps as facilitative tools in schools and corporations. Lawrence Erlbaum Associates, Mahwah, NJ (1998)

[12] Joseph. D. Novak (2006). "The Theory Underlying Concept Maps and How to Construct Them." Technical Report IHMC CMapTools 2006-01. Florida Institute for Human and Machine Cognition.

[13] Luiz Claudio Duarte Dalmolin, Silvia Modesto Nassar, Rogério Cid Bastos, Gustavo Pereira Mateus. (2009b) "A Concept Map Extractor tool for Teaching and Learning. Federal University of Santa Catarina (UFSC). 2009 Ninth IEEE International Conference on Advanced Learning Technologies.

[14] Luiz Claudio Duarte Dalmolin, Silvia Modesto Nassar, Rogério Cid Bastos (2009a) "Extrator de Conceitos e Termos Conectores para Criação de Cursos Baseados em Mapas Conceituais". Instituto de Informática e Estatística – Universidade Federal de Santa Catarina (UFSC)

[15] Miriam L. Domingues, Eloi L.Favero, Ivo P. Medeiros(2007) "Etiquetagem de Palavras para o Português do Brasil". TIL – V Workshop em tecnologia da informação e da linguagem humana. Rio de Janeiro (2007).

[16] Roy B. Clariana, Ravinder Koul, (2004) "A Computer-Based Approach for Translating Text into Concept Map-like Representations". The 1st International Conference on Concept Mapping. Pamplona, Spain.

[17] Shian Shyong Tseng, Pei-Shi Sue, Jun Ming Su, Jui-Feng Weng, and Wen-Nung Tsai, (2007). "A new approach for constructing the concept map". Educational. Computing. 49, 3 (Nov. 2007), 691-707.

Josiel C Reis received his B.Sc. degree in Civil Engineering from University of Amazonia in 1997. He specializes on Systems Engineering from ESAB (2012), and is presently a Computer Science MSc student at Federal University of Pará - UFPA, Brazil. Since 2010, he is a contributor to Brazil´s Government Teacher Training Program - PARFOR, on Computer Science Course Programs. He has experience on Web Systems Development and his main interests are on Text Mining, Software Process Improvements, Information Security, Distributed Systems and Real-Time Systems. He presently contributes for Institute of Education, Science and Technology of Pará, Brazil, on development and maintenance of Web Systems and Information Systems.

Sérgio Gaia received his BSc degree on Data Processing Technology from University Center of Pará – CESUPA (2003). He specializes in Data Base Systems from Federal University of Pará – UFPA (2007) and is a Computer Science MSc student at Federal University of Pará – UFPA, Brazil (2010). Since 2008 he has been contributing as an Information Technology Analyst for Institute of Education, Science and Technology of Pará, Brazil. He has experience in Computer Networks, Information Systems and Web Systems Development. His main interest areas are on Artificial Intelligence and Data Mining.

Raimundo Viégas Junior, received his BS degree on Electronic Engineering from Federal University of Pará – UFPA (1991) and holds a MSc degree on Telecommunication Engineering (UFPA-2002) and a PhD in Computer Engineering from Federal University of Rio Grande do Norte – UFRN (2010). He is presently a Professor at University Federal do Pará on graduate and undergraduate programs, and is also a contributor to the Secretariat of Logistics and Information Technology in State of Pará, Brazil. His research interests include Computer Networks Applied Computing, Wireless Networks and Real-Time Systems .