Extraction of Keywords from Texts: An Exploratory study using Noun Phrases

Renato Rocha Souza
Escola de Matemática Aplicada, Fundação Getúlio Vargas, Brasil. E-mail: renato.souza@fgv.br

K. S. Raghavan
Visiting Scientist. Centre for Knowledge Analytics & Ontological Engineering (KAnOE). PES Institute of Technology. E-mail: ksragav@hotmail.com

Abstract
The increasing use of Web for both scholarly publishing and information retrieval emphasizes the need for mechanisms to support efficient indexing and effective information retrieval. Manual indexing and knowledge representation techniques are not suitable for handling huge volumes of digital information. This paper presents an approach to extracting key phrases from texts based on the intrinsic semantics of the text. The methodology has been tested with a series of small-scale experiments involving texts in Portuguese language (SOUZA, 2005; SOUZA and RAGHAVAN, 2006). The results suggest that the approach yields satisfactory results. Some suggestions for future work have been made.
***Keywords*:** Information Retrieval Systems. Text Extraction. Text Semantics.

1 Introduction

The Web has emerged as a major player in information transfer and communication as a result of substantial increase in the volume of scholarly and business information published on the Web. This has underscored the need for more effective mechanisms for organization of information on the Web to facilitate efficient and effective retrieval than what is possible using the available search engines. Information and knowledge managers in organizations are increasingly facing a situation of both information overload on the one hand and increasing demand for filtered and relevant information on the other. Considering the sheer volume of information to be handled, it is generally conceded that solutions to handle such a situation should necessarily be technology-based and should make effective use of intelligent technologies. In fact the history of information retrieval clearly indicates that IR systems have always experimented with different strategies and new technologies to enhance information retrieval. Experiments aimed at using technology for enhancing information retrieval probably began with the early experiments by H.P. Luhn in keyword indexing and selective dissemination of information (SDI) in the middle of the 20th Century. Such efforts have continued to this day - directed at more efficient handling of information to facilitate and enhance retrieval and dissemination. Strategies that utilize digital computer technologies to manage large document collections have been in use for sometime now. Intranets including corporate portals, subject gateways and digital libraries are all developments along these lines. Developments and initiatives such as metadata initiatives (e.g. Dublin Core Metadata Initiative - DCMI), Semantic Web, ontologies, tools and technologies for data and text mining, etc

should all be viewed against this background. An examination of relevant literature suggests that the principal approaches that characterize the strategies in experimentation aimed at enhancing retrieval effectiveness include:

- How to improve the quality of metadata extracted from Web resources?
- What tools could be employed (e.g. Ontologies) to improve retrieval?
- How to design search interfaces that facilitate meaningful navigation in IR?

Keywords and key phrases present in a text are metadata of value in document representation and information retrieval. There are suggestions to the effect that identification and extraction of noun phrases (NPs), instead of keywords, may prove to be a more useful strategy for selection of index terms. This suggestion is based on the hypothesis that NPs carry the greater part of the semantics of a document, as opposed to articles, verbs, adjectives, adverbs and connectives (BAEZA-YATES; RIBEIRO-NETO, 1999, p.169-170). It has also been suggested that a substantial part of the semantics of the document, or the user query, could be lost when the text / query is represented merely by a Boolean combination of keywords. NPs have also been found to have other useful applications – e.g. for translation of concept maps (WOODS; RICHARDSON; FOX, 2005). There is also a certain amount of research that has specifically looked at the value and utility of NPs in the information retrieval context (Kuramoto, 1996 and 1999); Moreiro et al (2003); Velumani and Raghavan (2005 and 2006). *Parts-of-speech* tagging has been experimented with using texts in English language since the 1960s. Following these developments tools for handling texts in other languages – mainly the European languages – have been developed.

The principal objectives of the experiment reported in this paper are:

- To assess the utility and value of a methodology that has been developed to extract and assign scores (weights) to NPs in texts in Portuguese language; and
- To examine the value and utility of the top ranking NPs so extracted to serve as descriptors / key phrases to represent the '*aboutness*' of the documents from which the NPs were extracted.

2 Methodology

Souza and Raghavan (2006) have reported a work that describes a methodology for extracting appropriate NPs from texts in Portuguese language. The research essentially involved:

- Identification of NPs in texts using appropriate software agents;
- Developing a procedure for computing scores to be assigned to each NP, and assigning a score to every NP indicative of its value and utility as a key phrase in representing the '*aboutness*' of the document.

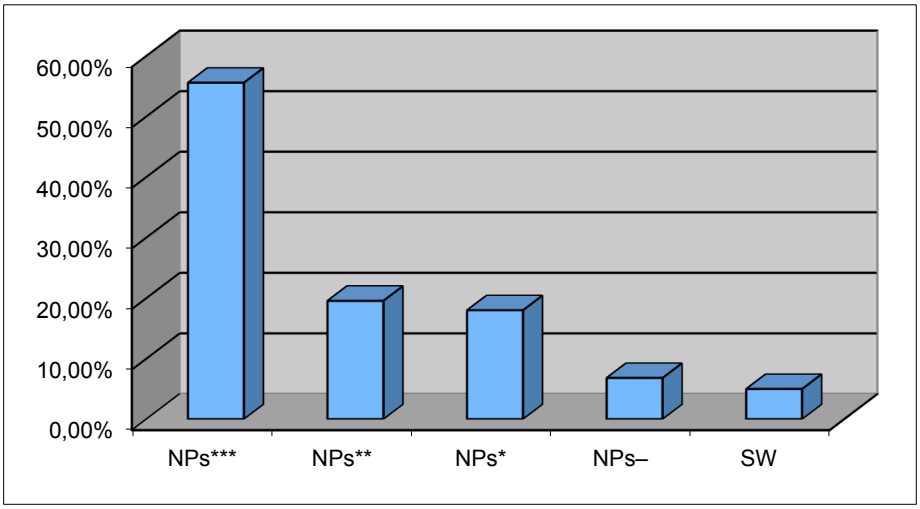
In this study three factors were considered for computing the Score (NP):

- Frequency of occurrence of a NP in a document (Tf);
- Inverse Document Frequency (IDf);
- A value called CNP – Category of Noun Phrase (based on the level & structure of a NP).

The score for every NP was computed using the formula, developed by the author (SOUZA, 2005):

$$Score\,(NP) = \left[\left(k1 * Tf\ (X)\right) - \left(k2 * IDf\ (Y)\right) + \left(k3 * CNP\right)\right]$$

Where, X, Y, k1, k2, k3 were all constants employed to correct distortions. The optimal values for X, Y, k1, k2, k3 were arrived at after manual examination and evaluation of the output. The final values suggested were those that yielded best results. The CNP values were also chosen by manual examination of data and ranged from 0.25 to 1. The methodology did yield very superior to methodologies that uses just plain keywords (SOUZA, 2005) in terms of identifying and extracting NPs representative of the '*aboutness*' of the documents. When both '*highly relevant*' and '*reasonably relevant*' NPs extracted were taken into account, about 70% of the NPs extracted were quite appropriate and could be considered good quality descriptors for the concerned documents. Figure 1 provides an overview of the results.



[Legend: *NPs**** = *Highly Relevant; NPs*** = *Reasonably Relevant*
***NPs** = *Moderately Relevant; NPs- = Not Relevant; SW = Stop Words*]**

Figure 1 - Overview of Output.

This research, when published (SOUZA; RAGHAVAN, 2006), received some comments essentially related to the '*arbitrariness*' of certain constants used in the method of assigning weights to NPs. A slightly revised methodology was developed and employed in further experiments and this paper reports the results of an experiment designed to incorporate some of these suggestions into the methodology and evaluating the output obtained using the revised methodology; it also tries to compare the output with the output of the earlier experiment. Figure 2 presents an overview of the methodology employed:

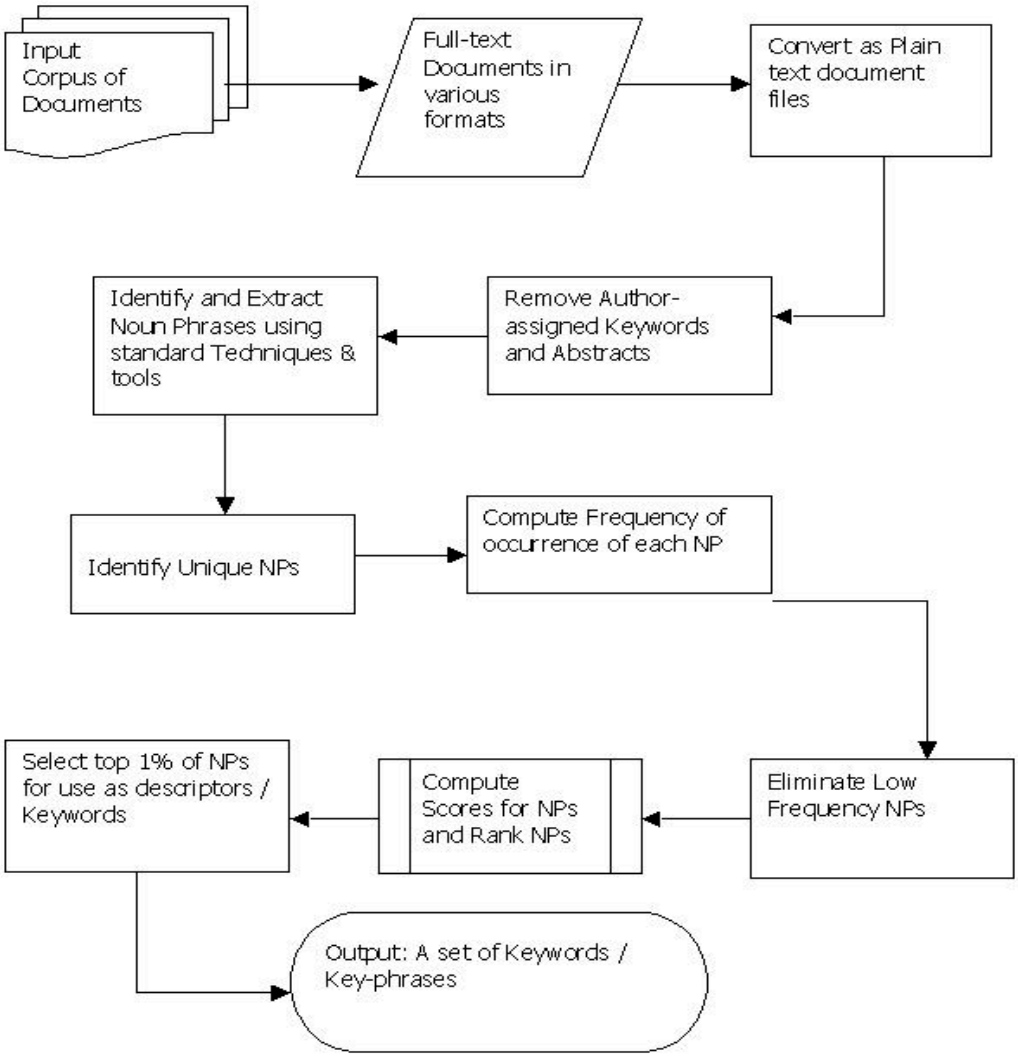


Figure 2 - Overview of Methodology.

3 Experiments and Analysis

The same corpus of 60 documents (SOUZA, 2005; SOUZA; RAGHAVAN, 2006) comprising of papers in periodicals in Portuguese language falling within the broad subject area of Information Studies that was the subject of the earlier study was experimented with using the revised methodology. Table 1 presents the details of the number of NPs extracted from each of the documents in the corpus.

Document #.	Total No of NPs	No. of Unique NPs	Document #	Total No of NPs	No. of Unique NPs
1	1673	1343	31	1702	1528
2	842	711	32	1902	1213
3	783	680	33	1941	1290
4	801	688	34	1480	1231
5	1478	1252	35	1011	788
6	984	836	36	735	552
7	638	521	37	2054	1382
8	779	684	38	772	624
9	1104	932	39	1873	1284
10	1146	1035	40	1156	962
11	619	554	41	1008	792
12	791	626	42	1244	1002
13	1342	1113	43	1808	1325
14	923	747	44	1375	1145
15	1063	877	45	1420	1176
16	888	810	46	1829	1453
17	1201	1084	47	987	810
18	5686	4287	48	1498	1223
19	1094	899	49	884	760
20	1299	1039	50	852	677
21	733	616	51	1225	1009
22	1837	1368	52	547	483
23	796	699	53	1364	1062
24	2048	1434	54	1535	1174
25	1368	988	55	1144	840
26	1246	1058	56	1386	1119
27	1173	971	57	1702	1353
28	788	667	58	1497	1166
29	617	539	59	733	632
30	633	506	60	1702	951
Average	1212.43	985.47	**Average**	1345.53	1033.53

Table 1- Number of NPs.

For most European languages the technology for identification and extraction of NPs from natural language texts is available. For concept extraction reasonably mature techniques (such as parts-of-speech tagging, word sense disambiguation, tokenizer, pattern matching, etc.) already exist and these have been employed in the field of information extraction. Suitable software packages for handling and processing texts in Portuguese language are available. 'PALAVRAS' (BICK, 2000), a parser developed at the Southern University of Denmark, and 'PALAVRAS XTRACTOR', developed jointly by the Universidade do Vale do Rio dos Sinos (Unisinos) in São Leopoldo, Brazil, and Universidade de Évora, Portugal were employed to process the NL texts and extract NPs.

The number of NPs (as well as the number of unique NPs) in any document is largely a function of the length of the text. On an average, each of the documents in the corpus contained about 1000 unique NPs (Ranging from a low of 506 NPs to a high of 4287 NPs). In principle every unique NP in a document qualifies to be considered for use as a descriptor (Keyword) indicating the '*aboutness*' of the document. However, in practice such an approach would result in employing a large number of descriptors for every document. While such a practice could contribute to substantially enhancing the recall output in a search, it is also known to result in unacceptably low levels of precision. Experience has shown that this indeed is the case with a large number of widely used Web search engines. It is therefore useful to identify the optimal level of exhaustivity of indexing. It was thought that about 1% of the unique NPs in a document would be adequate and optimal in terms of yielding acceptable levels of Recall and Precision. This was based on the fact that on an average the documents in the corpus had about 1000 unique NPs and 1% would yield about 10 descriptors per document. However, the level of exhaustivity can be changed if required.

It is common knowledge that many NPs found in a text have little value as descriptors. An important requirement, therefore, is to eliminate those NPs that are not likely to be of much value for use as descriptors in representing the '*aboutness*' of a document. A fairly simple and straightforward approach is to go by the frequency of occurrence of NPs in a given document. In this experiment also, all NPs occurring in a text with a frequency lower than a pre-determined threshold were ignored. Only those NPs, which occurred with a minimum level of frequency, were considered for further processing. This approach also helped in minimizing computational efforts required in the subsequent stages of the experiment. All the other extracted NPs were processed to compute their score and were ranked.

Several methods have been proposed and discussed for assigning weights to keywords / terms in a text. The most elementary method would be to assign a weight based on *term frequency*. A factor that affects the frequency of occurrence of a NP in a document is the length of the text. It is likely that a NP may occur with a greater frequency in a document that is longer than another document, which may also contain the same NP. Considering absolute frequency of occurrence of a NP for computing its score NP could, therefore, lead to distortions. Recognizing this, suggestions have been made to improve the methodology by using certain other data (e.g. inverse document frequency and normalized term frequency, etc). A couple of procedures were employed in order to obtain a comparative view of their effectiveness. The principal difference between what is reported here

and the earlier study relates to the methodology of computing the Score (NP).

3.1 The Results

Four different methods of assigning and computing weights of NPs [Score (NP)] occurring in the texts in the corpus were experimented with.

a) Computing weight based purely on normalized term frequency;

b) Computing weight based on both normalized term frequency and inverse document frequency;

c) Computing weight based on normalized term frequency, inverse document frequency and CNP factor;

d) Computing OKAPI[1] weight.

The procedure used to compute the weights assigned to NPs using the four methods mentioned above are given in the following table:

Method	Formula used for computing weight
Normalized term frequency	Score (NP) = f (i,j)
Normalized term frequency and inverse document frequency	Score (NP) = f (i,j) * log(N/n_i)
Normalized term frequency, inverse document frequency and CNP factor	Score (NP) = f (i,j) * log(N/n_i) * CNP
OKAPI	$\{Log\{N - ni + 0.5\}/(ni + 0.5)\} * 2.2 f(i, j)/\{0.3 + (0.9 * Ld / La) + f(i, j)\}$

[1] http://en.wikipedia.org/wiki/Okapi_BM25

Where:

- ***Score (NP*)** is the weight of NP i in respect of document j.
- ***f(i,j)*** is the normalized frequency of NP ***i*** in document ***j*** (The frequency is normalized to correct any distortion introduced by the document length; the normalized ***f(i,j)*** is got by dividing the absolute frequency of NP i in document j by the total number of NPs occurring in document j.
- ***N** is* the total number of documents in the corpus.
- n_i is the number of documents in the corpus which contain the NP i ***(log (N/ n_i)** gives the inverse* document frequency.
- L_d is the length of document **j**.
- L_a is the average length of documents in the corpus.

The Table 2 presents the number of unique and the number of NPs selected for each of the documents in the corpus (approximately 1% of the number of unique NPs extracted from the document).

Document #	No. of Unique NPs	No. of NPs selected	Document #	No. of Unique NPs	No. of NPs selected
1	1343	13	31	1528	15
2	711	7	32	1213	12
3	680	7	33	1290	13
4	688	7	34	1231	12
5	1252	13	35	788	8
6	836	8	36	552	6
7	521	5	37	1382	14
8	684	7	38	624	6
9	932	9	39	1284	13
10	1035	10	40	962	10
11	554	6	41	792	8
12	626	6	42	1002	10
13	1113	11	43	1325	13
14	747	7	44	1145	11
15	877	9	45	1176	12
16	810	8	46	1453	15
17	1084	11	47	810	8
18	4287	43	48	1223	12
19	899	9	49	760	8
20	1039	10	50	677	7
21	616	6	51	1009	10
22	1368	14	52	483	5
23	699	7	53	1062	11
24	1434	14	54	1174	12
25	988	10	55	840	8
26	1058	11	56	1119	11
27	971	10	57	1353	14
28	667	7	58	1166	12

29	539	5	59	632	6
30	506	5	60	951	10

Table 2: Number of NPs selected

For each of the documents in the corpus, all the unique NPs extracted were ranked on the basis of their weight and the top 1% of these NPs was considered as suitable to be used as descriptors. The selected NPs resulting from the different procedures employed for computing weights were manually examined for their value in terms of their appropriateness and suitability to be used as descriptors to represent the '*aboutness*' of the source document. The document in question was manually examined to assess the appropriateness of the NP. The quality of a NP was relative to whether the NP (or an equivalent term) figured in the set of keywords assigned by the author(s) of the paper.

The NPs were rated as NPs*** (Highly Relevant), NPs** (Reasonably Relevant), NPs* (Moderately Relevant) and NPs- (Not Relevant). Table 3 presents a comparative overview of the effectiveness of the different methods of assigning weights to NPs.

Method of Computing Weight → **NP Quality ↓**	Normalized Term Frequency	Normalized Term Frequency + Inverse Document Frequency	Normalized Term Frequency + Inverse Document Frequency + CNP	OKAPI Method
NP***	162	179	317	242
NP**	59	84	154	160
NP*	145	154	133	223
NP-	318	240	89	326
Total number	684	657	693	951

Table 3: Effectiveness of the different methods of assigning weights to NPs.

It was observed that the method based on the three factors, viz., normalized term frequency, inverse document frequency and CNP factor gave the best results in terms of its ability to yield good quality NPs. The results suggest that attaching weights based on frequency alone does not yield satisfactory results. The situation improved slightly when normalized frequency and inverse document frequency were employed to compute the weight. The results improved even further when CNP factor was also built into the computation process. Using OKAPI weights appeared to result in a large number of NPs being extracted. The main reason for this was that many of the NPs received the same scores. In identifying the top 10% of NPs to be used, all the NPs that had received the same scores were also considered. However, a comparison of the results obtained in these set of experiments with those obtained in the earlier research indicates that the previous methodology yielded even better results. One of the factors contributing to this could be the fact that a stop-word list was employed to suppress less important NPs in the experiments reported earlier. Since employing CNP factor yielded the best result, it was thought that it would be useful to experiment with varying the values of CNP to ascertain optimal values of CNP. The best results were achieved with the following CNP values.

Category	Structure and Level of NPs	CNP value
1a	Level 1, structure (D*+ N)	0.2
1b	Level 1, any structure except (D* + N)	0.8
2	Level 2, any structure	1.1
3	Level 3, any structure	1.4
4	Level 4, any structure	1.2
>4	Level 5 or higher, any structure	0.8

Table 4: Optimal cases / CNP values

4 Conclusions and Future Work

Any process for automatic keyword extraction could face two kinds of problems, which will affect the overall effectiveness of the system:

- Failure to recognize a semantically rich content-bearing term as such;
- Extracting a term of spurious and questionable value.

Failure to recognize an important term or phrase leads to an overall reduction in the recall performance of the retrieval system and extracting a NP of questionable relevance to the semantics of the document leads to a reduction in the precision performance of the system. Conceding the fact that manual cataloguing of Web resources is not a practicable solution, it is important for automatic extraction techniques to develop and adopt intelligent mechanisms for identifying and extracting NPs. The methodology proposed in this paper is a step in this direction. The methodology could be further refined and work along these lines is in progress. The principal approaches that are being considered in further refining the procedure are:

- To map the extracted NPs to descriptors in relevant vocabulary control devices such as thesauri / subject heading lists, etc. in an effort to both validate the NPs and, if possible, to standardize the terminology to the extent possible;
- To explore the feasibility and utility of generating and employing domain – specific lists of stop words (stop phrases) with a view to suppress generation of access points and retrieval under NPs of questionable value and utility;
- Document Expansion: In any process of extraction of NPs from a corpus of texts it is highly possible that some of the NPs that could have been there are not found in the text. This can and does affect retrieval effectiveness. In a slightly different but relevant context Singhal et al (1998 and 1999) proposed an approach that they have labeled as '*document expansion*' in an effort to enrich the documents in a collection with additional words *that could have been there*. It was reported that document expansion yielded substantial improvements in retrieval effectiveness. The technique can be effectively employed in IR research of the kind discussed in this paper. Applying document expansion techniques requires a source that can be used as the basis for expansion. The source could be in the form of a collection of texts or some other tools with similar topical content as the corpus of documents. It is not difficult to visualize a situation in which the extracted NPs could be expanded by adding synonyms and near synonyms of these taken from a thesaurus. This is an area that needs to be explored further.

A extração de palavras-chave a partir de textos: um estudo exploratório utilizando sintagmas

Resumo

O uso crescente da Web, tanto para a publicação acadêmica como para a recuperação de informação, enfatiza a necessidade de mecanismos para apoiar uma indexação eficiente e a recuperação eficaz da informação. Técnicas de indexação e representação do conhecimento manuais não são adequadas para lidar com grandes volumes de informação digital. Este artigo apresenta uma abordagem para a

extração de palavras-chave a partir de textos baseados na semântica intrínseca do texto. A metodologia foi testada com uma série de experimentos em pequena escala usando textos em língua Portuguesa (SOUZA, 2005; SOUZA e RAGHAVAN, 2006). Os resultados sugerem que o método produz resultados satisfatórios. Algumas sugestões para trabalhos futuros foram feitas.
Palavras-chave: Sistemas de Recuperação da Informação. Extração de Texto. Semântica do Texto.

References

Bick, E. **The Parsing System Palavras - Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework**. Aarhus: Aarhus University Press, 2000

Baeza-Yates, R. and Ribeiro-Neto, B. **Modern Information Retrieval**. ACM Press, New York:, 169-170, 1999.

Kuramoto, H. Uma abordagem alternativa para o tratamento e a recuperação de informação textual: os sintagmas nominais. **Ciência da Informação**, Brasília, v. 25, n. 2, 1996. Disponível em: <http://www.ibict.br/cionline/250296/25029605.pdf.>. Acesso em: 05 maio 2014.

Kuramoto, H. Proposition d'un Système de Recherche d'Information Assistée par Ordinateur Avec application à la langue portugaise. Tese (Doutorado em Ciências da Informação e da Comunicação) – Université Lumière - Lyon 2, Paris, France, 1999.

Moreiro, J. et al. Desarrollo de un Método para la Creación de Mapas Conceptuales. Anais do ENANCIB, Belo Horizonte, 2003.

Singhal, Amit et al. ATT at TREC-7. In The Seventh Text REtrieval Conference, 239-252, November 1998. Disponível em: <http://trec.nist.gov>. Acesso em: 05 maio 2014.

Singhal, Amit. and Pereira, Fernando. Document expansion for speech retrieval. In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 34-41. ACM Press, August 1999.

Souza, R. R.; RAGHAVAN, K. S. A methodology for noun phrase-based automatic indexing. **Knowledge Organization**, v. 33, n. 1, p. 45-56, 2006.

Souza, R. R. Uma proposta de metodologia para escolha automática de descritores utilizando sintagmas nominais. Doctoral Thesis. UFMG, Belo Horizonte, 2005.

Velumani, G. and Raghavan, K. S. Automatic Extraction of Keywords from Web resources. **Information Studies**, 11(3): 185-194, 2005.

Velumani, G. and Raghavan, K. S.. Extraction of Keywords: A Noun Phrase-based Methodology (In Knowledge Representation and Information Retrieval edited by K. S. Raghavan. – Bangalore: DRTC, Indian Statistical Institute, paper P, 2006.

Woods, John O. Richardson Ryan & Fox, Edward A. Multilingual Noun Phrase Extraction Using a Part-of-Speech Tagger, 2005. Disponível em:<http://www.writing.eng.vt.edu/Abstract/John%20Woods.pdf >. Acesso em: Acesso em: 10 maio 2014.