Language Independent System for Document Context Extraction

Jiri Pribil, Tomas Kincl, Vladislav Bina, Michal Novak

***Abstract*—At this time all people, especially managers and businessmen, are exposed to the ever-present information pollution. This is why tools of business intelligence are of great importance; nevertheless the current methods can hardly cope with large and unstructured text sources like World Wide Web that currently becomes more and more important. To achieve this main goal we have to find and verify satisfactorily reliable methods for automatic extraction of a main context of a document, i.e., multidimensional structured characterization representing the main topic of the document. To cope with the multilingual sources we have to develop approaches that would not be dependent on the language of the source and that would not need any additional language dependent tools (like thesauri). In our conception, the context is dynamic – it means that a classification of a document will not be dependent only on the document in question but also on the corpus; the expansion of a corpus can result in a change of a document classification.**

***Index Terms*—business intelligence, context extraction, text mining, unstructured information sources**

I. Introduction and Motivation

The Internet has already become a serious source of information and has an important role among corporate PR activities. Traditional mass media are no longer the main channel used for publishing information.

More and more Internet users are not just web content consumers, but they are also becoming authors of user generated content. A further increase of importance of the Internet is expected in connection with the fact that already in 2013 experts anticipate that more than 60 % of all web users will become active producers of web content. Such user generated content is spreading among customers to support their buying behavior and stresses the importance of web monitoring for an everyday life of a successful manager.

Manuscript received July 15, 2011; revised August 11, 2011.

This work was supported in part by the Czech Republic National Research Programme II under project No. 2C06019 "(Medical) Knowledge Acquisition and Modelling" and by Prague University of Economics Internal Grant Agency under project No. F6/6/2011 "Automated methods of context extracting from unstructured text based on the fundamental statistical and linguistic models".

Jiri Pribil is an Assistant Professor and Vice Dean for External Affairs at the University of Economics, Prague, Faculty of Management, Jindrichuv Hradec, Czech Republic (e-mail: pribil@fm.vse.cz).

Tomas Kincl is lecturer and researcher the University of Economics, Prague, Faculty of Management, Jindrichuv Hradec, Czech Republic (kincl@fm.vse.cz).

Vladislav Bina and Michal Novak are PhD students at the University of Economics, Prague, Faculty of Management, Jindrichuv Hradec, Czech Republic (e-mail: bina@fm.vse.cz, novak@fm.vse.cz).

To some extent successful monitoring systems have already been implemented. They usually consist of two parts: a crawler-robot, which searches the web for the specified data, and an analytic interface, which "interprets" meaning of the data found.

While the first part of such a system, i.e., a crawler-robot, is often efficiently implemented, the latter analytic part is usually a weak part of the system. Imagine the goal is to find documents bearing a price of a required product. It is obvious that a relevant document need not bear the word "price" but terms like "expensive", "cheap" or "cost". In reality, the situation is even more complicated because it may happen that the required information is published in another language. This explains why these services are at present mostly provided by human analyst and there is a huge demand for such specialists on the job market.

Nevertheless, unstructured text processing can be realized with the help of already published methods and algorithms but usually it requires some additional information about the used language – especially dictionaries of synonyms. Our research aims at multilingual and meta-data independent methods for unstructured text processing that will take all the necessary additional information from a corpus.

This paper introduces first results of the system based on our research. To illustrate the results we present some experiments using free *American National Corpus Second Release - Open Portion (Open ANC)* [1], part *written_2\travel_guides*. This test corpus contains a total of 179 papers written in the English language with the average length of 36,981 characters.

A. Current state of Work

For example, the initial research on the field of the automatic classification of customer complaints has been made [2]. The methods of text analysis, however, always require additional linguistic tools and are closely linked to the specific language [3, 4]. Our system is designed as a language independent and self-learning.

B. Practical Application of the Research

The aim of the research is subsequent application in practice. The above mentioned usability for company managers (monitoring of customers' opinions, following the opinion about the competitors, etc.) seems to be interesting.

Quite different area of interesting application could be the automated documents cataloging – every company produces and receives a huge amount of electronic material and this data must be stored in a smart way (the documents should be properly "tagged" for later use.)

ISBN: 978-988-18210-9-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

II. Definitions

Our research is based on the use of two key elements that need to be defined – context and unordered *n*-gram.

A. Context

As a context of a document we understand its description (representation) enabling a quick and simple identification of the content of the document and allowing some additional operations on the documents like their comparison (based on the similarity of their contexts) or their subsequent aggregation into "thematic groups". In fact, there are many ways how the context may be represented. Let us mention the most important, or rather most often used, approaches (the list is ordered according to the increasing complexity of the representation):

i. simple list of significant words;
ii. ordered list of significant words according to their importance;
iii. probability distribution on a set of significant words (weighted list);
iv. hierarchical structure on a set of significant words (tree);

The *context* should not be confused with the concept of a *keyword*; there are three major differences:

i. the keyword must be manually defined (or at least checked after automatic suggestion) for each source document (document must be "evaluated" by hand),
ii. keyword must appear in the document, while the words in context need not necessarily occur in the document,
iii. unlike the keywords, there is no predetermined list of words in the context; they are generated dynamically and are able to respond to emerging trends in the texts.

The context does not have a predetermined number of elements - the document can only have a strong background (and thus a relatively simple context) or may be characterized by more topics (and thus a larger context). The method of calculation and representation of document context is one of the main subjects of this paper.

B. n-grams and {n}-grams

The basic term in the field of text processing is that of an *n*-gram. The classical concept defines *n*-gram as a *n*-element vector consisting of *n* adjacent words (or multi-word terms) in the text (Fig. 1).

Original text:

Lorem ipsum dolor sit amet, quis nostrud sit dolor ipsum elit.

Text representation:

2-gram	Count	Frequency (%)
lorem ipsum	1	10
ipsum dolor	1	10
dolor sit	1	10
sit amet	1	10
amet quis	1	10
quis nostrud	1	10
nostrud sit	1	10
sit dolor	1	10
dolor ipsum	1	10
ipsum elit	1	10

Fig. 1. Sample text and its representation using classic 2-grams (standard preprocessing applied).

A huge shift in this field, which was introduced in the recent work of our team [5] is an application of unordered *n*-grams. Unordered *n*-gram – {*n*}-gram – is defined as an *n*-element set consisting of *n* adjacent words in the text. Two different *n*-grams that differ only in the ordering of the words are thus represented by the same unordered *n*-gram (Fig. 2).

Original text:

Lorem ipsum dolor sit amet, quis nostrud sit dolor ipsum elit.

Text representation:

{2}-gram	Count	Frequency (%)
ipsum lorem	1	10
dolor ipsum	2	20
dolor sit	2	20
amet sit	1	10
amet quis	1	10
nostrud quis	1	10
nostrud sit	1	10
ipsum elit	1	10

Fig. 2. Sample text and its representation using unordered {2}-grams (standard preprocessing applied, {2}-grams are created as alphabetically sorted 2-grams).

This increases, for example, flexibility of algorithms for an intelligent automatic text comparison and analysis. In our previous research, this concept was used in a wide range of cases, especially in advanced plagiarism detection [6].

In document context extraction system we use the unordered *n*-grams for both the detection of stop words (more precisely stop terms composed of *n*-words) and the actual extraction of context.

III. Premises

Focused on the expected application area (business intelligence) we express the following basic assumptions and requirements:

i. there is a corpus of documents (document storage, document warehouse), into which the processed documents will be stored;
ii. the system can process any text obtained from any source (company document server, customers or business partners business documents, contributions from the online discussions, social networks, chats);
iii. new documents entering into the system do not go through any human control or preprocessing (see the only exclusion later) and can be written in any language using alphabetic writing systems (not logographic or syllabic writing systems), especially using the Latin alphabet;
iv. new documents entering the system are written in a language which is presented in the corpus;
v. the system automatically detects the language of the document and performs the analysis and extraction of context based on the information from the preprocessed corpus;
vi. the context will be dynamic - it means that the classification of a document will not be dependent only on the document in question but also on the corpus; an expansion of a corpus can result in a change of the context.

ISBN: 978-988-18210-9-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

IV. Document Analysis

The following description of a context extraction procedure reflects the current state of knowledge of the team and current implementation of methods tested in various practical experiments:

A. each document entering the system goes through the basic stages of preprocessing - at least formatting and punctuation removing,
B. the document is transformed into {*n*}-grams,
C. based on the {*n*}-grams analysis, the language of the document is recognized,
D. stop words ("information insignificant" words) are removed from the document,
E. the document is stored in the corpus along with the information about its language, both in full text and as {*n*}-grams representing the document,
F. document context is extracted based on basic statistical characteristics of specified words or {*n*}-grams and their relationship to other {*n*}-grams both in the corpus and the analyzed document.
G. document context is visualized for human inspection or for more precise specification of system parameters.

Particular methods and their parameters, which are mainly used in the last two steps, are the main subject of the research and practical experiments in this project.

A. Document Linearization

Document linearization is a process of basic document content filtration. There are usually two steps – (a) *markup and format removal* (all markup tags and special formatting are removed from the document and the document is converted to plain text) and (b) *tokenization* (all remaining text is lowercased and all punctuation is removed as well as the number sequences; thus, the document is represented as one very long sentence).

B. n-grams Transformation

We are using {*n*}-grams rather than classic *n*-grams. This alternative abstracts from the order of words and is used together with the stop words removal with the goal of even further information concentration of the document content.

C. Language Identification

The process of language identification is theoretically described [7, 8] and uses mainly statistical approaches. In most cases, the document is represented as a set of classical *n*-grams, but these *n*-grams are created as a subsequence of letters (character *n*-grams) instead of words. In our system, we use comparison of dynamically generated stop words (for each language) to automatically detect the language of the document.

It is obvious that in the case of empty or only partially filled corpus the system will not be able to correctly identify the document language. In that case, it is necessary to perform the "human classification" – the results of our experiments show that the required number of such manually recognized documents is very small; count in the order of ones for longer documents and at most in the order of tens for short documents in similar languages (for example Russian/Ukrainian or Portuguese/Spanish).

D. Stop Words Removal (Stop Listing)

The first problem in the process of text analysis is the categorization of words. As the stop words we recognize commonly utilized words which are not important for the content of document. In the classical concept of stop words there is a list of these defined for each language – and every author and system uses significantly different list [9]. These stop words must be defined manually – and that's not very useful in the language independent system.

We automatically distinguish words as stop words in a simple way: stop words are the most utilized words in each language (words with the highest frequency in the corpus for given language). In praxis, the threshold can be defined either absolutely (for example as a word with the occurrence of more than 3 ‰ in the whole corpus) or relatively (for example top 100 words form the corpus).

Fig. 3 shows the results of our experiments on the above mentioned test corpus. Although we (theoretically) don't know anything about the language at all, the list really contains words (the most common words throughout the corpus) without significant "information value". So, we can describe these words as the stop words.

Total No. of words: 1,022,952
Total No. of unique words: 43,355

Stop words analysis:

No.	Word	Count	Frequency (‰)
1	the	88,425	86.44
2	of	39,396	38.51
3	and	33,241	32.50
4	a	24,002	23.46
5	in	21,012	20.54
6	to	20,583	20.12
7	is	13,521	13.22
8	s	9,728	9.51
9	for	8,353	8.17
10	on	7,923	7.75
11	with	7,338	7.17
12	from	7,077	6.92
13	are	6,484	6.34
14	by	6,215	6.08
15	as	6,154	6.02
16	it	5,930	5.80
17	you	5,905	5.77
18	at	5,783	5.65
19	was	5,171	5.05

Fig. 3: Stop words analysis results (corpus: *Open ANC*, part *written_2\travel_guides*, standard preprocessing applied). Only stop words with frequency > 5 ‰ are displayed. Another 74 words have frequencies above 1 ‰ (*that, its, an, but, or, this, has, be, one, most, can, which, city, century, th, their, have, town, there, de, here, more, were, also, all, his, many, where, see, they, some, old, not, into, island, museum, up, out, other, only, over, than, around, been, two, who, new, first, when, along, will, north, built, world, small, just, after, still, south, street, now, if, place, years, km, through, great, your, miles, road, park, church, well, center*).

A similar analysis can then be made for *n*-grams or unordered {*n*}-grams. In this case, the results of the analysis of *stop terms* (instead of *stop words*) are even stronger in terms of the subsequent analysis of valuable context. Fig. 4 shows the example of 2-grams analysis.

ISBN: 978-988-18210-9-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Total No. of 2-grams: 963,563
Total No. of unique 2-grams: 411,897

Stop words analysis:

No.	Word	Count	Frequency (‰)
1	of the	12,610	13.09
2	in the	6,355	6.60
3	to the	4,603	4.78
4	on the	3,067	3.18
5	and the	2,957	3.07
6	from the	2,500	2.59
7	is the	2,479	2.57
8	th century	1,946	2.02

Fig. 4: Stop terms (2-grams) analysis result (corpus: *Open ANC*, part *written_2\travel_guides*, standard preprocessing applied). Only stop terms with frequency > 2 ‰ are displayed. Another twelve terms have frequency above 1 ‰ (*at the, is a, for the, by the, the city, one of, the th, with the, of a, and a, as the, you can*).

The main disadvantage of language independent algorithm is undoubtedly the fact that the algorithm is not able to deal with various forms of the word. That leads especially in the case of Slavic languages to the fact that the same stop word is expressed by more inflexions (tense, grammatical mood, grammatical voice, aspect, person, number, gender, case). There are also some fragments ("nonwords") as the result of document preprocessing and numbers or punctuation removal ("s" and "th" in Fig. 3 and Fig. 4 above). Practical experiment shows, however, that this is not a problem.

E. Documents Storing

Any modern relational database can be used to store the corpus. Most operations are then performed by optimized SQL queries. At this point, we use the PostgreSQL database because of its best performance.

F. Document Context

The core of the system – analysis of the document context – is based on a premise similar to the analysis of stop words. For the purposes of this section we define *language corpus* as the part of the corpus in the same language as the analyzed document.

In the most simple scenario, the individual elements of context are defined as words or $\{n\}$-grams (or generally *terms*), which satisfy two conditions:

i. their frequency in the *document* is relatively high,
ii. their frequency in the corpus is relatively low.

For each word or $\{n\}$-gram in the analyzed document that was not identified as a stop word we calculate:

$$\varphi(t_i) = fd(t_i) / fc(t_i). \quad (1)$$

Here t_i is i-th term (word or $\{n\}$-gram) in analyzed document, $fd(t_i)$ is the frequency of a term t_i in analyzed document, and $fc(t_i)$ is the frequency of a term t_i in the corpus. Because terms t_i from analyzed document are stored in corpus just before this step, the value of $fc(t_i) > 0$ and thus the value of $\varphi(t_i) > 0$. The upper bound of $\varphi(t_i)$ is not limited.

As with stop words, the elements of context can be defined both absolutely – the context is composed of the terms with the value of $\varphi(t_i)$ higher than the selected threshold – and relatively – the context is composed of the selected number of terms with the highest values of $\varphi(t_i)$.

Document context will be represented as a set of terms t_i ordered by descending value of $\varphi(t_i)$ – see Table 1 for three examples.

TABLE 1
DOCUMENTS CONTEXT ANALYSIS RESULTS

Document	Context Words	ϕ
HistoryJapan.txt	*shoguns*	66.900419
	tokugawa	54.892651
	edo	46.252141
	samurai	46.252141
	authority	39.644692
	shinto	35.680223
	japan	34.839275
	nara	29.042042
	threat	28.929911
	japanese	28.104285
Nepal-WhereToGo.txt	*stupa*	42.071397
	nepal	34.520121
	tibetan	31.031385
	shiva	24.477904
	pilgrims	12.605662
	buddhist	7.761599
	temple	7.746172
	stands	7.397169
	valley	7.311284
	golden	6.437154
WhatToMadeira.txt	*wicker*	118.979035
	madeira	84.810214
	funchal	76.437113
	quinta	58.993772
	rua	58.878324
	monte	43.759116
	holes	40.223026
	embroidery	39.504758
	items	35.428471
	serra	26.336505

Corpus: *Open ANC*, part *written_2\travel_guides*). Context in absolute form – 10 words with highest φ selected, *fd* > 0.001.

Another improvement of document context calculation can be achieved by including some "term weights" in the $\varphi(t_i)$ formula. The introduction of the *similarity of words* could be another huge move in the process of improving the context calculation methods and could compensate the role of lemmatization in our language independent system. This concept, which is algorithmically and computationally much more demanding, is in the phase of testing.

G. Visualisation

The visualization of the extracted document context is very important especially in the first phase of system development and fine-tuning.

Fig. 5 shows a graphical analysis of the document context using heat maps. In this view, the document is represented by individual cells, each cell as a word or multiword term in document. The darkness of the cell corresponds to the importance (strength) of the context – the darker the cell, the stronger the context; identical words are represented, of course, by the same shade.

According to the threshold set for relatively calculated context terms φ, the map (Fig. 6) changes respectively as the number of terms added to the context varies. This allows us to improve the currently tested algorithms

ISBN: 978-988-18210-9-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

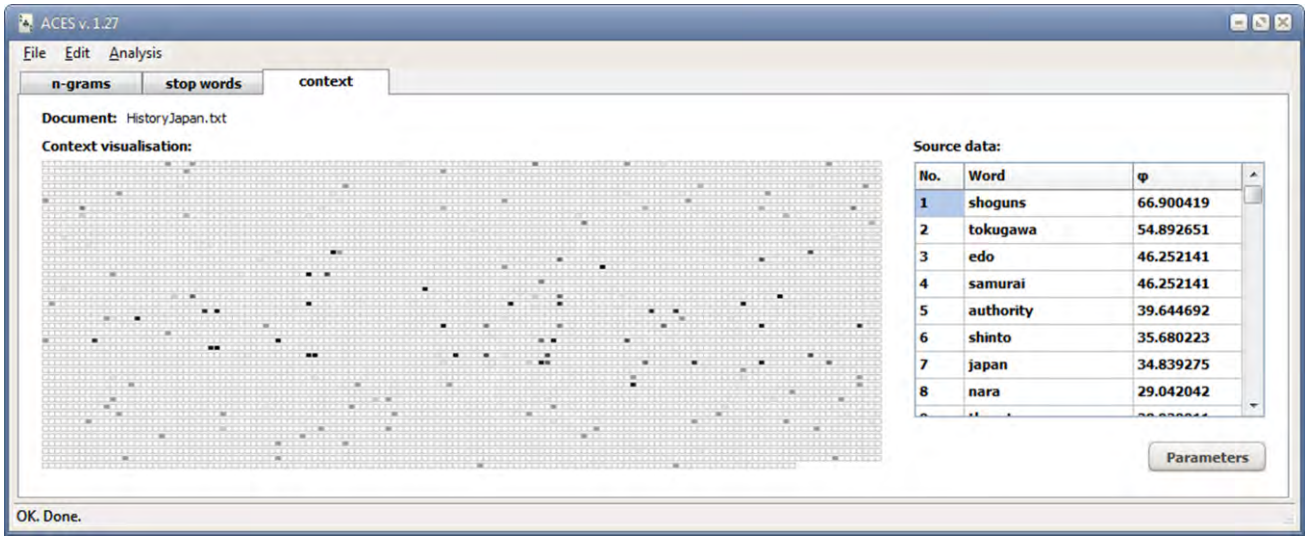

Fig. 5: Document context analysis results (corpus: *Open ANC*, part *written_2\travel_guides*, analyzed document: *HistoryJapan.txt*). Context in relative form (totally 10 terms with $\varphi > 20$ selected).

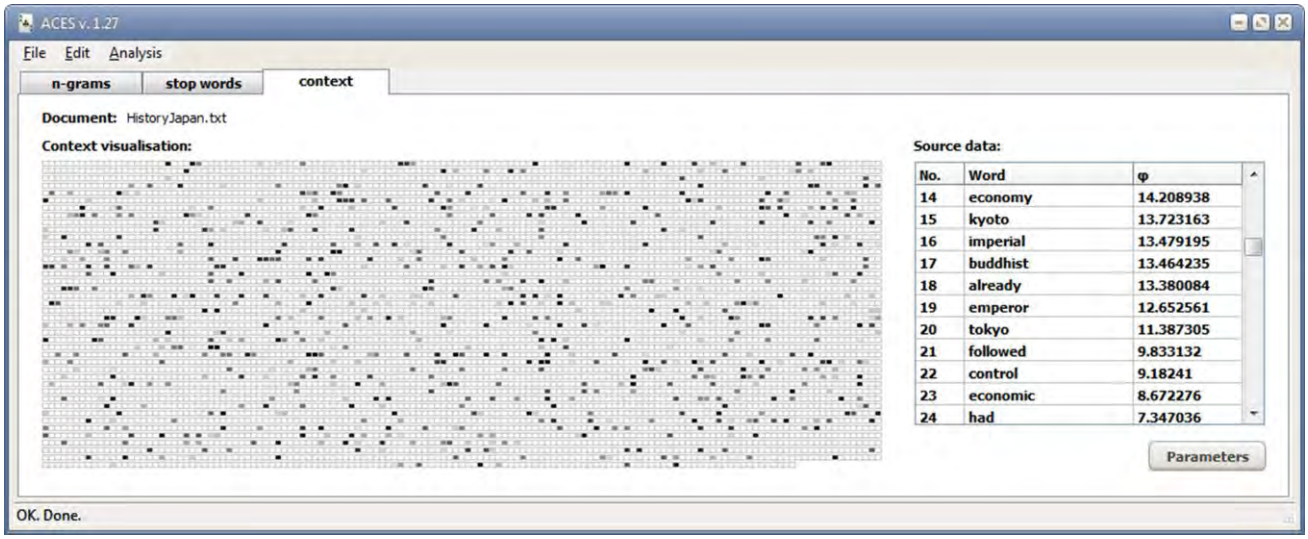

Fig. 6: Document context analysis results (corpus: *Open ANC*, part *written_2\travel_guides*, analyzed document: *HistoryJapan.txt*). Context in relative form (totally 53 terms with $\varphi > 2$).

V. Future Work and Conclusion

Our research achieved some interesting and encouraging results both on the field of document context extraction and visualization.

At this moment we are working very hard on selecting more appropriate description of context instead of simple list of significant words presented in this paper.

Further work will also aim to improve the visualization of the results – particularly different graphs (network outlining the relationships between elements of the context, tree graphs) seem to be very promising.

References

[1] *American National Corpus Second Release - Open Portion* Linguistic Data Consortium, University of Pennsylvania. Available: http://americannationalcorpus.org/OANC/

[2] Coussement, K. and D. V. d. Poel, "Improving customer complaint management by automatic email classification using linguistic style features as predictors". *Decis. Support Syst.*, 2008. vol. 44, no. 4, pp. 870–882.

[3] Allan, J., "Introduction to topic detection and tracking", in *Topic detection and tracking 2002*, Kluwer Academic Publishers, pp. 1–16.

[4] Nallapati, R., "Semantic language models for topic detection and tracking", in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, vol. *3*2003, Association for Computational Linguistics: Edmonton, Canada. pp. 1–6.

[5] Pribil, J. and H. Kalinova, "Types of text plagiarism detectable by unoriented n-grams", in *13th Czech-Japan Seminar on Data Analysis and Decision Making in Service Sciences* University Hall: Otaru, Japan, pp. 211–218, October 2010.

[6] Pribil, J., O. Leseticky, and H. Kalinova. "Plagiarism at Universities – How to Fight It? Case of the Czech Republic", in *International Conference on Information Communication Technologies in Education*. Kerkyra, Greece, July 2010, pp. 122–133.

[7] Baldwin, T. and M. Lui, "Language identification: the long and the short of the matter", in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*2010, Association for Computational Linguistics: Los Angeles, California. pp. 229–237.

[8] Prager, J. M., "Linguini: Language Identification for Multilingual Documents", in *Proceedings of the Thirty-Second Annual Hawaii International Conference on System Sciences, vol. 2*, IEEE Computer Society. pp. 20–35.

[9] Brahaj, A., "List Of English Stop Words". Available from: http://armandbrahaj.blog.al/2009/04/14/list-of-english-stop-words/.

ISBN: 978-988-18210-9-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)