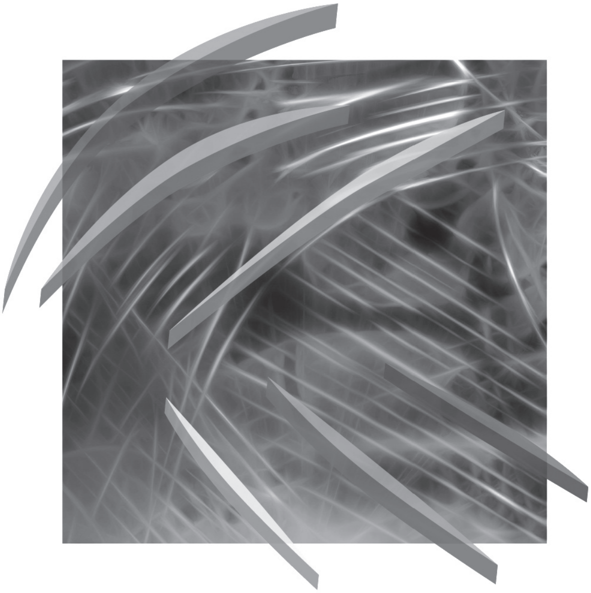
INFORMATYKA EKONOMICZNA

BUSINESS INFORMATICS

21 • 2011



Publishing House of Wrocław University of Economics
Wrocław 2011

Copy-editing: Agnieszka Flasińska, Elżbieta Macauley, Tim Macauley,

Layout: Barbara Łopusiewicz

Proof-reading: Barbara Cibis

Typesetting: Małgorzata Czupryńska

Cover design: Beata Dębska

This publication is available at www.ibuk.pl

Abstracts of published papers are available in the international database The Central European Journal of Social Sciences and Humanities http://cejsh.icm.edu.pl
and in The Central and Eastern European Online Library www.ceeol.com

Information of submitting and reviewing papers is available on the Publishing House's website
www.wydawnictwo.ue.wroc.pl

ISSN 1507-3858 (Business Informatics)
ISSN 1899-3192 (Research Papers of Wrocław University of Economics)

The original version: printed

Printing: Printing House TOTEM
Print run: 200 copies

Contents

Streszczenia

INFORMATYKA EKONOMICZNA BUSINESS INFORMATICS 21 • 2011

ISSN 1507-3858

Anna Filipczyk
University of Economics in Katowice, Katowice, Poland
anna.filipczyk@ue.katowice.pl

USING TEXTUAL STATISTICS TO SUPPORT COMPETITIVENESS COMPANY ANALYSIS

Abstract: Knowledge of the internal and external activity aspects of an enterprise is its essential resource. Companies communicate on-line with others through their own websites, which allows for easier identification of a company and finding new potential clients. Knowledge of business opponents gives a company the opportunity to follow market trends and to compete with other enterprises. Additionally, it may verify the attractiveness of its products and services in relation to businesses competing in a similar market by analysing various websites. To make this review faster and more automatic, a company can use an application based on the statistical text analysis method, which gives an opportunity to detect similarities between websites and therefore contributes to their competitiveness and company competitiveness in general.

Key words: knowledge, internet, semantic analysis, text, website.

1. Introduction

The most valuable resource of a company is its knowledge, influencing both value and success of a business, an incorporeal property owned and made by people, and used to determine directions of development (actions). Its intangibility and instability along with high difficulty of evaluation differentiate knowledge from other company resources. Furthermore, collecting knowledge, though being a highly time consuming activity, is in fact one of the few resources which increase during exploitation. Knowledge resources owned by an enterprise apply to its internal activity aspects, such as information on products offered, technologies used or employers' experience, as well as external aspects, e.g. market knowledge, future clients or rivals.

For every company, information contained in plain text documents is a valuable source of knowledge resulting from an efficient analysis of such content, as it influences the quality of a firm's services and products. The nature of knowledge and its growing significance for companies have caused a rise in popularity of the term "knowledge management" and in the use of systems employing such techniques.

2. Internet as a resource of the knowledge of competitors

The wide use of the Internet and the growth of its meaning as a basic information search tool are undeniable. Almost 88% of companies have their own websites and more than 60% use the Internet in order to establish their business contacts. Small businesses as well as international concerns use the Internet to communicate with their clients through their own websites. Nowadays, owning a site is not only an element of prestige and high reputation of a firm but additionally a necessity and an obligation that increases clients perception. Because of a website a company can be easily identified, it may gain greater possibilities of establishing contacts with clients and demonstrate its activity to a wider range of recipients. However, the Internet is not the only means of presenting an activity profile.

Internet technology provided companies with yet another range of possibilities. Assuming a website is a basic promotion tool for a business, analyzing web resources widens knowledge about potential competitors. The market position of a company ought to be analysed as the first element, followed by an identification of competitive organisations. The latter can be found on the net as firms with similar products and services. Through the analysis of the ways of their development and their basic activities searching through the Internet allows a company to gain knowledge about its rivals as well as to research the competitiveness of their own website.

A company's success is determined by the quality of its products and services, yet ignoring competitive businesses on the market prevents it from noticing possible threats, giving a delusive sense of safety instead. The recognition of the range of competitors' capabilities is substantial in increasing the effectiveness and efficiency of a company. Additionally, the success of an enterprise depends on its ability to adapt to continuously changing surroundings. As moderations of the market are unavoidable, it is the innovative technique, organisation and management that allows a company to become a leader in its industrial grouping. A company aware of the competitive firms' activities is more likely to precede its rivals by its own actions or to improve its weaknesses. Knowledge of potential threats from competitors may be helpful in case they enter the market, as it allows to recognize the market situation of a particular business, the popularity of the offered products and the services and type of management.

The problematic aspects of identifying the market situation and competition indicate the need for a company to employ an IT solution as an effective tool not only for finding business opponents' websites, but additionally for supporting competitiveness level analysis of its range in relation to potential market rivals.

3. Latent Semantic Analysis in measuring the similarity of documents

Text mining is based on an analysis of plain text documents by removing phrases, terms, etc., and preparing the processed text in a way enabling further numerical analysis. Statistical text analysis is a text-exploring method related to the quantitative approach of quality analysis of written text, inspired by an idea and a dream of creating a computer system or program capable of examining text in a human manner.

The first stage of plain text document analysis is building a dictionary corpus, that is a set of words to be studied, which begins with the identification of a lexicometric base made of a number of graphic elements; a graphic element can be defined as a set of characters, usually letters, surrounded by punctuation and strictly connected with a particular word. In the case of a website analysis, a lexicometric base of corpus is the content of a website.

The creation of a text corpus dictionary is followed by a lemmatisation. This is a technique which allows for connecting graphic elements with one particular word in different forms, all deriving from this one base from, e.g. one form of a word is a verb used in a particular form, noun singular or plural. Words are grouped according to particular parts of speech derived from the same word core. However, the technique of lemmatisation includes some drawbacks, such as a danger of connecting words with the same base form but with different meanings. Lemmatisation uses a data base which includes words in their basic form and changed forms. A basic word is identified for each word depending on its location in a proper table of data base.

Another step in text analysis is digitisation, a procedure consisting of interposition of corpus on graphic elements and of numbering or addressing each form. During further analysis, forms are kept in the corpus dictionary, whereas addresses or numbers of graphic elements are used for research purposes.

Tables of frequency are built after the process of digitisation. Such a table consists of data on the frequency of occurrence of a particular word and aims to eliminate words with too low frequency.

Another step of this scheme is creating a table of distance. In some methods of statistical text analysis making a distance table is meant to divide the corpus into specific classes. In such a case, table rows correspond with lemmated words, and table columns – with categories or groups of analysis purpose. Values 1 or 0 are introduced into table cells depending on whether the answer includes a word or not; a balance of particular words due to the whole text may be used instead. Further text analysis consists of statistical methods; there can be distinguished corpus answer classification, vector representation of documents or analysis of hidden semantic groups of documents. In the presented example the method used was Latent Semantic Analysis (LSA).

Latent Semantic Analysis is applied for representing the meaning of words and larger parts of a text, such as sentences and paragraphs. The LSA method is connected with artificial neuron networks, still, it depends on matrix decomposition according to Singular Value Decomposition (SVD), the mathematical technique of reduction of matrix dimension, with the help of which texts, including in their content knowledge close to the particular subject, from the point of view of volume to the knowledge owned by experts of particular subject, are studied. Thanks to LSA procedure it is possible to simulate human appraisal of affinity of words meaning or text fragments.

The LSA method depends on the use of the reduction frequency according to peculiar values [Lula 2005, pp. 8–14] and corresponding with the matrix.

Figure 1 illustrates the Singular Value Decomposition. The output element of the analysis is the creation of the matrix **X**, which is disposed due to peculiar values. Matrix columns match the analysed; respectively, rows correspond with particular words used for documents similarity level study. The matrix **X** elements are reflections of the frequency of word occurrence in a particular document. These can be natural numbers which characterize the amount of occurrence or weight of a particular word in a document (expressed as a fraction). Creating the matrix **X** is strictly connected to the described lemmatization process, as the frequency of occurrence of particular words (or weights) is calculated in relation to the search in the content, for the word in different forms. In the approach of frequency of occurrence as words weights, such values are calculated by products of the local weight and global weight. The use of global functions allows to decrease the most frequent words weight. One of the methods of calculating local weights is to use a logarithm (a quotient of occurrence amount of the word in a document, plus 1). On the other hand, to calculate global weights a normalizing function may be used (assignation of an inverse element from the sum of square roots of local functions). The use of normalising function allows for applying greater weight to words less frequent in analysed texts, however, the use of logarithmic function allows for the elimination of big numbers from the calculations. The use of weights for words is the most frequent and the most efficient method of perfecting the outcomes of the described text analysis method.

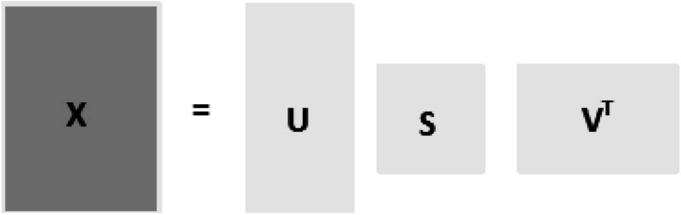


Figure 1. Singular Value Matrix Decomposition

Source: [Lula 2005, p. 13].

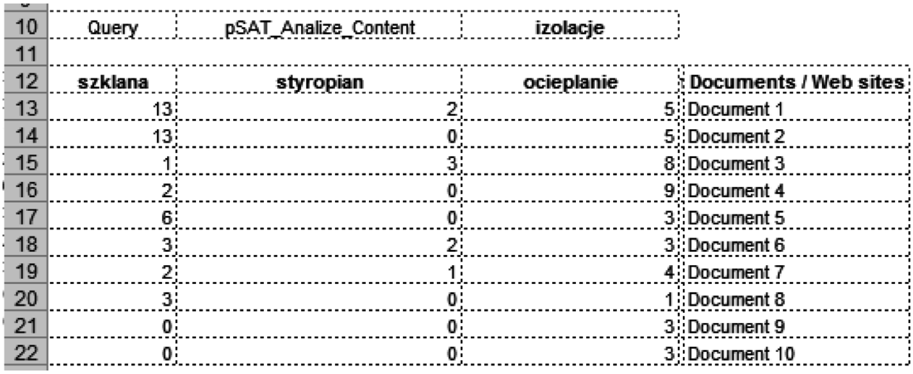
The purpose of further calculations in this process is to define space in which the analysis of the sets of words appearing in the documents (matrix **U**) and the analysis of document sets (matrix **V**) are possible. Each object (word, document) is represented by one row of matching matrix, but particular coordinates are arranged decreasingly; due to its informative value, such order allows for space reduction by considering only some number of the initial vector elements. The classification of the documents depends on the reflection of this scheme to the matrix **V**, where its elements scaled by the matrix **S** data determine points representing analysed documents.

The level of the similarity of documents considered with the help of the scalar product of the matrix rows is achieved by the multiplication of matrix $\mathbf{V}^{\mathbf{T}'} \times \mathbf{S}'$. An empirical study showed that connection of analysed texts is presented by a number from the range of (0; 1.5). Such numerical range evaluating compatibility of text results from the use of the weight function approach. The outcome of the analysis, close to 1.5, suggests a high similarity between documents; on the other hand, the outcome close to 0 shows a low connection between the content studied.

4. The experiment

The problems presented above, along with the problematic aspects of company competitiveness analysis indicate the extreme importance for a company to search for particular businesses and to analyse their activity profiles. Such type of system creates an opportunity to expand a company's knowledge of its rivals and of their activity profiles, which allows it to notice competitors' weak and strong points.

Figure 2 presents a part of an application main module implemented in MS Excel. The application was made for analysis purposes in Polish. The results of a documents search with the use of a key word *izolacja* (Polish for "insulation") are visible in the picture (addresses of web pages are hidden). Parameters for analysis are *szklana* (Polish for "glass"), *wata* ("wool"), *styropian* ("extruded polystyrene foam"). The combination of words was prepared in order to find producers of insulation materials.



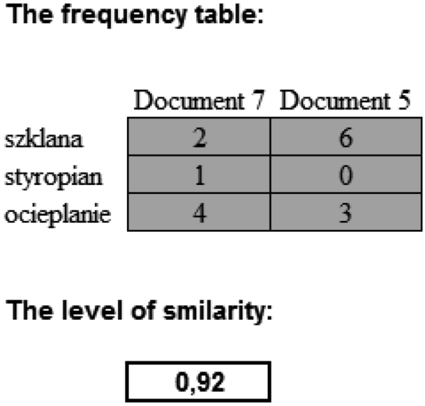
10	Query	pSAT_Analize_Content	izolacje	
11				
12	**szklana**	**styropian**	**ocieplanie**	**Documents / Web sites**
13	13	2	5	Document 1
14	13	0	5	Document 2
15	1	3	8	Document 3
16	2	0	9	Document 4
17	6	0	3	Document 5
18	3	2	3	Document 6
19	2	1	4	Document 7
20	3	0	1	Document 8
21	0	0	3	Document 9
22	0	0	3	Document 10

Figure 2. Search documents result

Source: own elaboration.

The search result is in the form of a list of links to websites of companies offering building insulation services. The higher number of parameters may indicate the greater probability of getting results more interesting for the user.

In order to analyse the level of the similarity of documents, the user ought to choose from among the websites found. Firstly, document 7 (website 7) and document 5 were chosen as an example. The conducted latent semantic analysis showed the level of the similarity at 0.92, which means an average level of similarity. Figure 3 shows given words frequency table and the result of the analysis.



The frequency table:

	Document 7	Document 5
szklana	2	6
styropian	1	0
ocieplanie	4	3

The level of smilarity:

0,92

Figure 3. Similarity analysis result

Source: own elaboration.

In the second order document 7 and document 3 were analysed. The comparison produced a result of 1.33 similarity level, which indicates greater similarity than was received in the first analysis. According to the results of these two studies, the company whose range is presented on the website 3 (document 3) is most probably more similar to the range of company 7 than to the range of company 5.

The use of the LSA method for gaining knowledge of competitive companies may facilitate even an on-line search for potential business partners. The Internet is a source of valuable information and is becoming an increasingly important source of data. However, the growing amount of information put on the net and its chaotic nature may result in growing difficulty of finding truly interesting and significant data and using it as valuable knowledge. Finding important firms, from competitors' and potential business partners' point of view, may prove to be highly laborious and time consuming. The tool described above is employed to make the process of finding particular profile companies more efficient. The system discussed allows for an introductory analysis of activity range convergence of searched companies by studying the level of content similarity on websites, using the discussed text analysis method (LSA).

The system presented makes a quick search through websites and a study of their content possible. Moreover, owing to the use of statistical text analysis, the application facilitates comparing documents in order to find businesses most similar to the model.

5. Conclusions

Knowing an enterprise's competitor may be a significant factor determining its success. The Internet gives them both an unlimited possibility independently of whether they look for clients or want to present their activity profile. However, searching for a particular piece of information on the net is highly difficult and time consuming due to the great amount of data. Therefore, the automation of on-line data searching is of great importance, especially for company employees responsible for the surrounding analysis and the level of company competitiveness study.

The use of statistical text analysis in collecting knowledge of a company competitiveness is considerably profitable. It allows for the efficient processing of competitors' data and indicates those most similar to the model, which can be the website content of the enterprise analysing its market position. The tool presented allows for searching websites for future clients and, additionally, for conducting an introductory analysis of the similarities of a company profile activity with the aim of beginning cooperation. Thus, employment of a tool based on statistical text analysis may have a profound contribution to a company's development, eliminating its weak points and improving its strong points, and therefore leading to its more efficient functioning and to greater profitability.

References

Damerau F., Indurkhya N., Weiss S., Zhang T., *Text Mining. Predictive Methods for Analyzing Unstructured Information*, Springer, New York 2005.

Deerwester S., Dumais S.T., Furnas G.W., Harshman R., Landauer T.K., *Using Latent Semantic Analysis to Improve Access to Textual Information*, ACM, New York 1988.

Gołuchowski J., *Technologie informatyczne w zarządzaniu wiedzą w organizacji*, Wydawnictwo Akademii Ekonomicznej, Katowice 2007.

Lula P., *Text Mining jako narzędzie pozyskiwania informacji z dokumentów tekstowych*, StatSoft Polska, 2005, http://www.statsoft.pl/czytelnia/8_2007/Lula05.pdf.

Manning C., Schutze H., *Foundations of Statistical Natural Language Processing*, Massachusetts Institute of Technology, Cambridge, MA, 1999.

Mykowiecka A., *Inżynieria lingwistyczna. Komputerowe przetwarzanie tekstów w języku naturalnym*, Wydawnictwo PJWSTK, Warszawa 2007.

Rabino G., Scarlatti F., *Textual Statistics, Conceptual Mapping, Bayesian Networks and Landscape Evaluation*, 2002, http://www.ersa.org/ersaconfs/ersa02/cd-rom/papers/483.pdf.

Siwek A., Kowalska M., Szczyt M., Statystyczna analiza tekstu (Textual Statistics), [in:] E. Soja, A. Ptak-Chmielewska, A. Siwek, M. Rodzewicz (Eds.), *Nowe metodologiczne propozycje analiz w naukach społecznych ze szczególnym uwzględnieniem demografii*, Sekcja Analiz Demograficznych KND, Warszawa 2000, pp. 52–66.

ZASTOSOWANIE STATYSTYCZNEJ ANALIZY TEKSTU DO WSPOMAGANIA ANALIZY KONKURENCYJNOŚCI FIRMY

Streszczenie: Wiedza dotycząca aspektów wewnętrznych i zewnętrznych działań firmy jest istotnym zasobem. Zarówno małe, jak i międzynarodowe korporacje komunikują się bezpośrednio z innymi poprzez własne strony, które pozwalają na łatwiejszą identyfikację firmy i znalezienie nowego potencjalnego klienta. Z drugiej strony Internet jest także bardzo wartościowym źródłem informacji o możliwych klientach, partnerach biznesowych czy konkurentach. Wiedza dotycząca biznesowych konkurentów nie tylko daje możliwość naśladowania trendów rynkowych i konkurowania z innymi firmami, ale także identyfikuje pozycję na rynku. Dodatkowo można zweryfikować atrakcyjność oferty w relacji do biznesowych konkurentów na podobnym rynku poprzez analizę różnych stron internetowych. Aby szybciej i w sposób bardziej zautomatyzowany dokonać takiej oceny, firma może zastosować aplikację, opierając się na statystycznej metodzie analizy tekstu, co daje możliwość wykrycia podobieństw pomiędzy stronami internetowymi. Na podstawie zdobytej w ten sposób wiedzy organizacja ma możliwość zwiększenia konkurencyjności swojej oferty, a w konsekwencji polepszenia swojej pozycji na rynku.

Słowa kluczowe: wiedza, Internet, analiza semantyczna, tekst, strona internetowa.