KNOWLEDGE SEARCH IN INTERNAL AND EXTERNAL DOCUMENTS*

L. Dömeová, M. Houška

Czech University of Agriculture, Faculty of Economics and Management, Department of Operational and Systems Analysis, Prague, Czech Republic

The documents which are not in possession and under control of user cannot be changed according to the users needs. On the other hand this group of texts is numerous and can be processed using various text mining tools. The results of automatic knowledge mining are in many cases not much beneficial. After all processing of a great number of documents of different authors on the same topic can bring new implicit knowledge. In the article we propose a combined method for realizing important characteristics of a product from documentation published by different rival firms. The internal documents of firms are usually designed for complete reading by subject person. Different situation comes over in situation of need of quick response. Such situation can be connected with rapid changes in marked conditions, actions of competitors. In such cases there it is necessary to find quickly proper documents or their parts. We propose to analyze and reshape the documents using soft system methodology. According to the knowledge flow analysis the structure of documents can be designed or changed. The header of documents should be completed with corresponding keywords which can be used for automatic or semi automatic knowledge search.

structured and unstructured documents; document analysis; quick response; knowledge flow; keywords

INTRODUCTION

Recently business decision making under uncertain environment of target markets becomes more difficult. To judge investment adequately, it is indispensable to analyze various factors, for instance, not only financial aspects but also aspects of organizational growth, customers satisfaction and so forth (Loebbecke, Wareham, 2003).

For decision making the managers need to draw information and knowledge from various documents. A huge number of potentially useful texts are placed on Internet and in external and internal databases which are often highly substantial.

The problem of insufficient information was replaced by information overloading. No human decision maker is able to process all available documents, it means to collect, to evaluate their relevance and reliability, to study them rigorously and apply the information or knowledge gained in practical situation properly. On the other hand, along with the development of managerial information systems, it became possible to use some artificial tools as data warehouse or data-mining (Brožová, Havlíček, 2005).

Another serious problem is lack of time for majority of decisions. Rapid and intensive development and exploitation of communication technique led to shortening the period for decision making from days to hours and minutes (Šubrt, 2004). Globalization and other aspect typical for world market environment made the strategic decisions highly important, unrepeatable and irreversible. The expression "one shot decision" depicts typical features of this decision making which can be compared from the point of view of time pressure and extreme importance with necessary fast reactions in emergency situations as fire or spate.

The strategic scenario, which consists of logical actions for continuously producing enterprise outcome, plays a significant role in making management along with business strategy. It takes much time to construct strategic scenario by newly emerging information only by hand of human professionals. Therefore there is a strong need to make it possible to construct strategic scenario by information technology, which is expected to support human decision makers (Masanori Akiyohi, Norihisa Komoda, 2004).

The main reasons for text analysis and extracting data are then an informational overloading and lack of time of each decision making person. Complications are also connected with complexity, heterogeneity and multidimensionality of the problems solved. For solution of such complicated tasks there is generally no acceptable body of knowledge such as rules, laws, etc., which may assist professionals. Rather, knowledge lies in memories of experiences of past performance and one must understand past decisions (own as well as other peoples') to make better future choices (Reich, Kapeliuk, 2005). While there are more approaches, we focused on document analysis based decision support.

The organizations and their people vary significantly. Organizations have different strategic goals and value systems, structures and practices. The people involved

* The paper was supported by a grant project of the Ministry of Education of the Czech Republic No. MSM6046070904 – "Information and Knowledge Support of Strategic Management".

may also come from different culture and value backgrounds and that affects their working and communication styles.

Therefore, it has long been recognized that there is no single development method for information systems (McFarlan, 1981). Similarly there is no single method for text analysis. The big shortcoming of well-known mining model such as the vector space model (Salton, 1971) and some statistical models (Katz, 1995). is that the performance of text mining is not satisfactory even for limited text documents in many situations. Other techniques may come to have difficulties with the large number of available documents (Lean et al., 2005).

We tried to analyze two types of documents. In the article we call them "structured" and "unstructured" but it really does not mean that the documents have or have not any inherent structure. The classification is connected with the ability of user to affect the structure. It may be possible to call them also "influence able" and "non-influence able".

The first group we call "unstructured documents" which means the structure and content of them is out of control of the user. These documents are usually placed in the Internet or in some public or private databases. The user has no owner's or author's rights to change anything in them. As example we can mention public market-related documents such as collection of consumer opinion, research results of potential needs for products or services, commercial descriptions of products of competitors, offers and experience of customers in e-shops, etc. There might be thousands of documents condescended with one type (class) of product. The knowledge obtained from these documents for example about consumers' needs and opinions seems to be very general and not much reliable. We think that the reliability grows with raising number of texts analyzed. The analysis of hundreds of documents from different sources can finally bring valuable knowledge which is not in fact explicitly mentioned in single documents and that's why it cannot be obtained by studying only several pieces. The analysis of high number of texts is impossible without some automatized or semi-automatized support.

In a following case study we propose a combined method for searching typical features of products using characteristics made by different producers.

The second type of documents we call "structured" what means that the user is able to create or change the content or structure of the documents. In this group we focused on internal documents which contain rules, former solutions and experiences which can be used in situation when quick response is needed. Each firm, region and state prepares and maintains a set of documents which can be used for solving emergency situations. In commercial enterprises the emergency situation is not necessarily connected with natural disasters but also with technical accidents, extremely rapid and crucial changes in the market, unrespectable behavior of competitors, criminal offence, etc.

The pre-prepared documents are the sources of knowledge for decision makers in these situations. However, the documents are released by different departments and concerned with different aspects of the knowledge. Therefore, it is difficult to find and extract all related knowledge pieces from the documents quickly (Rong, 2005). We suppose that common structure of these documents can be helpful both for the decision maker and also for the operational subject (employees and operational departments).

MATERIAL AND METHODS

Analysis of external documents

We here explain our approach on extracting data on significant features of competitive products. The analysis can find which product properties must be emphasized in new product development as well it can help in evaluation of market position of already existing product. The most important features of product common for competitive producers can be used as criteria for multiple criteria evaluation.

The proposed approach has following steps:

1. Collection of appropriate text

   The easiest way is to use some web search engine. For a general question one can get thousands of matching texts what is an advantage in this method.
2. Finding the most frequent words in texts

   This finding can be made by some commercial text mining tools (http://www-306.ibm.com/software/data/iminer; http://www.megaputer.com).
3. Removing all words which occur in less than certain percent of documents

   The percentage can be set adequately to concrete situation.
4. Choosing typical noun phrases

   This step should be made by human expert. The expert extracts the noun phrases that are considered to reflect the product concept from the set of most frequent words. He/she may also find and join together synonyms so that they will be further treated as the same expression. These phrases will serve as description patterns.
5. Finding words matching the description pattern

   The words matched the description pattern are extracted as keywords which reflect the main features of the product.

The procedure with description patterns should avoid choice of words which do not really describe the product. The most frequent words have usually higher grammatical than meaning importance.

CASE STUDY

Finding considerable features of LCD monitors

For demonstration of above described process we collected 20 texts (from 108 000 000 offered by the web

Table 1. Typical phrases and keywords in LCD monitors descriptions

Phrase	Occurred times	Keywords	Occurred times
Provides, supports, secures	26	Connector	17
It is possible to connect	17	Point size	16
Users call for, pay attention	14	Native resolution	12
		High resolution	12
		Response time	11
		Ergonomics	10
		High quality of picture	10
		OSD menu	9
		Convenience	9
		Maximal security	8
		Easy control	8
		Modern design	6
		Acceptable price	6
		Small size and weight	5
		High contrast	5

search engine) on LCD monitors produced by 7 various companies (http://www.eizoshop.cz). After removing all words which were in less than 1/4 of documents and joining synonyms together into one expression we found 3 typical phrases, which we further used for assigning keywords. After it we searched for words matching these phrases. Many of these keywords and phrases were found in majority of documents, see Table 1, even several times in one text. Using these keywords and phrases we are able to describe important features of the product and make mutual comparisons of monitors of different producers.

The described method is simple and the results of the case study were obtained on a small sample without using any professional text mining tool. Even thought the results can be used as a guide for potential customer or professional for setting evaluation criteria. The most frequently mentioned properties can be considered as the most important ones. We think that in this particular case bigger sample will not bring deciding changes into the results.

Generally more advanced use of the Internet, i.e. changeover from passive access to distributed information to creation and exploitation of implicit is desirable. However, there is not an integrated and comprehensive modeling environment to efficiently utilize the knowledge resources available on the Internet (M a k o w s k i , 2005).

EXTRACTION OF KNOWLEDGE FOR QUICK RESPONSE

We concentrated on internal firm documents which are used in situation which demands of quick response. The internal documents are primarily designed for being read completely by target subjects. The authors of these documents do not suppose that the documents will be worked up by artificial tools.

The situation is generally different in emergency, when only some knowledge pieces are needed. The emergency situation for company is not necessarily only natural disasters, terror attacks or industrial accidents but also some fast and crucial changes in economical and marketing environment which can cause serious economic loss. Quick response in such terms is the core approach to reduce the loss. On the other hand the situation with quick answer demand could not represent only potential threat but also unrepeatable chance.

Although there are many differences among emergencies according to R o n g (2005) it is always necessary to provide the knowledge about know-what, know-how, know-when, know-which and know-where for the decision maker to enable him make quick response. When a problem occurs the subject and functional departments will deal with the objects. The functional departments should execute the task with related resources such as manpower, materials, and funds and so on.

The emergency response can be also described by task sets and an object sets (Fig. 1).

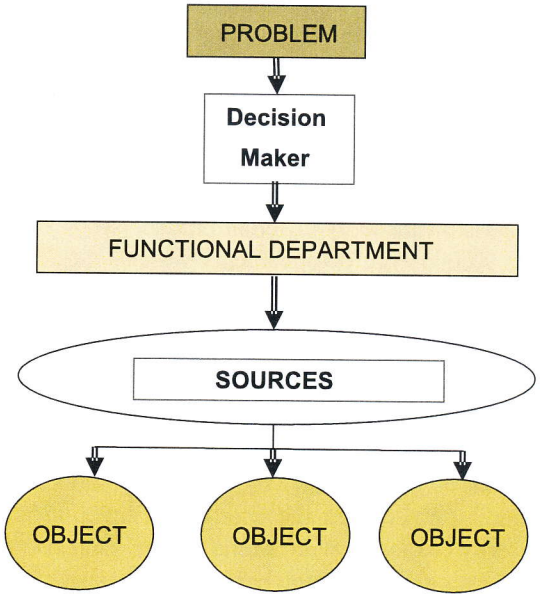

Fig. 1. Problem-driven quick response

Knowledge requirement for decision maker

When problem occurs the decision maker has to determine who will deal (the department, team, and person) with the problem – functional department. For this decision it is necessary to know *what, when* and *where* it happened. Know *how* implies knowing the solution of the problem. The problem can be divided into several objects which may become tasks for various functional departments (Fig. 2).

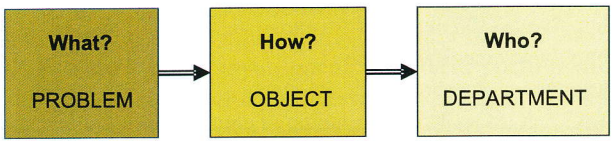

Fig. 2. The knowledge flow for decision maker

Knowledge requirement for functional departments

The functional departments execute certain tasks. So they need to have the knowledge about the task solution and about the sources necessary including their location. (Fig. 3)

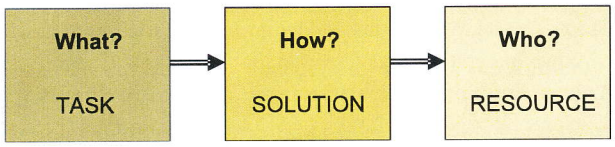

Fig. 3. The knowledge flow for functional department

Quick response as transformation process

For the definition of the documents structure we can use soft system methodology which describes each activity as a transformation process in which some "input" is changed, or transformed, into some new form of the same entity, the "output" (C h e c k l a n d , 1981).

S m y t h and C h e c k l a n d (1986) suggested defining (root definition) each purposeful activity in consideration of elements CATWOE. These elements can be also used for definition of above mentioned points of the knowledge flows, i.e. *what*, *who* and *how*, see Table 2. The quick response can be understood as a transformation process which input is a "need for decision" and output a "reaction to emergency situation". In the framework of the CATWOE definition a need for three more specifications appears:

Owner – those who could start and stop the transformation (in emergency situation this person must be uniquely determined)

Customer – those who would benefit or will be damaged by the decision (own firm, rivals, specific group of employees)

Weltanschauung – makes the activities meaningful (goals and policies, firm culture).

Basic structure of documents

The main problem is that the knowledge needed is spread in a number of various documents and it is hard to get required knowledge pieces in time. The goal of reorganization of the firm documents is to support the knowledge flow from decision maker to the functional departments and afterwards execution of tasks by finding proper solutions and assignment of sources.

The reorganization of the firm documents should lead to standard structure with clear rules concerning allocation of documents and access rights. The documents for quick response are generally of two types:

a) Experience
- Past solutions and their results
- Proven decision making processes
- Best practices

b) Directives
- Emergency plans
- Internal rules and regulations
- Law and rules established by state or local authorities comprehend into internal documents

Table 2. The CATWOE elements and knowledge flows

C	Customers		The victim or beneficiaries of the decision
A	actors	*Who?*	who would realize the decision DEPARTMENT RESOURCE
T	transformation process	*How?*	how the problem or task will be solved OBJECT SOLUTION
W	weltanschauung		general goals, firm policies, context, world view
O	owner		decision maker
E	environmental constraints	*What?*	description of the situation PROBLEM TASK

The document for quick response in fact contains description of activities which should be carried on in certain situations. These activities are transformations from "need for decision" to "reaction" or "response". Matching the problem situation with appropriate knowledge pieces follows the knowledge flows for decision maker or/and for the functional department (see Figs. 2 and 3). Each transformation can be matched with knowledge of *what, how* and *who* as well as with the *owner, customer* and *context.*

Hence the content of the documents should shortly and simply describe the above mentioned 6 points. The structure of the documents should be reorganized so that the knowledge of above mentioned types can be found quickly, if possible by means of some artificial tools.

RESULTS

Knowledge has become one of the most important factors for organizations. The rapid increase of available texts raises the importance of methods which can effectively find useful knowledge for support of human decision makers suffering from informational overloading.

In the paper we tried to compare methods of gaining knowledge from two different types of documents.

First group contains huge amount of documents of different structure and content what is typical for texts on the Internet. Exploitation of these texts seems to aim only to shallow and inaccurate results. But when we take advantage of great number of available texts we may obtain useful knowledge which is not explicit in them. In searching for typical property of a product we found that the same expressions occur in majority of texts of different producers. From this fact we derived that these properties are important and should be observed and compared. We recommend using these properties also as criteria for multiple criteria decision making.

Second group consists of firm documents which can be used in the situation of quick response needed. Internal documents are usually designed for being read whole and properly by human subjects. In emergency situation separate pieces of knowledge must be found and used rapidly. The investigation of knowledge needs of the subject (decision maker) and the object (functional departments) of decision making followed the knowledge flows. Both the subject and the object need answers to questions what, how and who. Answers to these questions should format the basic structure of described documents. From the point of view of the activity as a transformation process there should be three more characteristics: the owner, the customer and the weltanschauung (worldview, context). Unique structure of documents produced by different departments or authors will be a support for quick and correct decision and task execution.

Even thought extracting the knowledge from different types of documents might not be easy and various tools and methods need to be designed; exploitation of all possible sources of knowledge is groundwork for a sustainable competitive advantage.

REFERENCES

BROŽOVÁ, H. – HAVLÍČEK, J.: Data, informace a znalosti v matematických modelech. In: Proc. Conf. Agrarian Perspectives, Praha, 2005: 1080–1089.

CHECKLAND, P. B.: System Thinking, System Practice. Chicester, John Wiley 1981.

http://www-306.ibm.com/software/data/iminer

http://www.megaputer.com

http://www.eizoshop.cz

KATZ, S. M.:(1995) Distribution of content words and phrases in text and language modeling. Natural Language Engineering, *2*, 1995 (1): 15–59.

LEAN YU – SHOUYANG WANG – KIN KEUNG LAI: Multi/agent based Web Text Mining on the Grid for Crude Oil Price Prediction. Int. J. Knowledge System Sci., *2*, 2005 (2): 39–52.

LOEBBECKE, C. – WAREHAM, J.: The impact of eBussiness and the information society on "STRATEGY" and "STRATEGIC Planning": An assessment of new concepts and challenges. J. Inform. Technol. Management, *4*, 2003: 165–182.

MAKOWSKI, M.: Virtual Modeling Laboratories for Knowledge Integration and Creation. In: Proc. Int. Congr. The New Roles of Systems Science for a Knowledge-based Society, Kobe, Japan, 2005.

MASANORI AKIYOHI – NORIHISA KOMODA: An Analysis Framework of Enterprise Documents for Strategic Scenario/based management. Graduate School of Information Science and Technology, Osaka University, Japan, 2004.

MCFARLAN, F. W.: Portfolio approach to information systems. Harvard Business Review, *59*, 1981 (5): 142–150.

REICH, Y. – KAPELIUK, A.: A framework for organizing the space for decision problems with application to solving subjective, context/dependent problems. J. Decision Support Systems, *41*, 2005: 1–19.

RONG, L.: A method of managing the knowledge in government documents for quick response. Int. J. Knowledge System Sci., *2*, 2005 (2): 67–73.

SALTON, G.: The SMART Retrieval System – Experiments in Automatic Document Processing. Englewood Cliffs, New Jersey, Prentice Hall, Inc. Publishing 1971.

SMYTH, D. S. – CHECKLAND, P. B.: Using a system approach: the structure of root definitions. J. App. System Analysis, *5*, 1986 (1): 75–83

ŠUBRT, T.: Multiple Criteria Network Models for Project Management. Agric. Econ., *50*, 2004 (2): 71–76.

Received for publication on June 12, 2006

Accepted for publication on August 21, 2006

DÖMEOVÁ, L. – HOUŠKA, M. (Česká zemědělská univerzita, Fakulta provozně ekonomická, katedra operační a systémové analýzy, Praha, Česká republika):

Vyhledávání znalostí v interních a externích dokumentech.

Scientia Agric. Bohem., *37*, 2006, Special Issue: 38–43.

Významný rozdíl při získávání znalostí z dokumentů představuje faktor vlastnictví dokumentů, ze kterého vyplývá možnost ovlivnit jejich strukturu. Dokumenty jsme z hlediska možné analýzy a získávání znalostí rozdělili na interní a externí.

Externí dokumenty nejsou ve vlastnictví firmy nebo osoby, která je používá, tzn. není možno do nich zasahovat a měnit jejich strukturu. Jedná se zejména o veřejné materiály, zprávy publikované konkurenčními firmami, vědecké podklady, zákony a další regulace státních a místních orgánů.

V příspěvku navrhujeme kombinovaný postup pro získání důležitých charakteristik vybraného produktu z propagačních materiálů různých firem. Postup je založen na analýze textu z hlediska vyhledávání nejfrekventovanějších slov (podstatných jmen) ve velkém počtu textů. Tato slova mají význam, pokud jsou ve spojení s určitými slovesy a pokud se vyskytují ve velkém procentu zkoumaných textů. Tímto postupem se nám podařilo vyhledat určité charakteristiky výrobků, které jsou důležité a mohou sloužit např. jako kritéria pro vícekriteriální rozhodování. Tato metoda dává pouze přibližné výsledky, její význam se zvětšuje při velkém počtu textů.

Do druhé skupiny patří dokumenty, které vytváříme nebo jejichž strukturu jsme schopni upravovat. To jsou interní dokumenty prvotně určené ke čtení a studiu pro příslušné zaměstnance a předpokládá se, že budou přečteny podrobně celé. Nejsou tedy většinou připraveny pro jakékoliv automatizované zpracování. Výjimku představují situace vyžadující rychlou odpověď. Typické příležitosti vyžadující rychlou odpověď na státní nebo regionální úrovni jsou nejrůznější krizové situace, jako jsou přírodní katastrofy a válečné konflikty. V prostředí firem patří do situací vyžadujících rychlou odezvu nejen podobné krajní situace (požár, průmyslová havárie), ale i rychlé změny trhu a akce konkurentů, které mohou představovat jednak ohrožení a jednak i příležitost, kdy při rychlém jednání je možno výhodně nakoupit, získat konkurenční výhodu apod. Struktura interních dokumentů by proto měla být připravená pro rychlé vyhledávání dokumentů i jejich částí.

Pro tento účel je vhodné definovat znalostní potřeby subjektů i objektů rozhodování pomocí toků znalostí. Třebaže se jednotlivé dokumenty i situace značně liší, vždy je třeba pro správné rozhodnutí získat znalosti typu CO? KDO? JAK? (R o n g , 2005). Pro definici struktury dokumentu je pak možno použít měkkou systémovou metodologii, která chápe každou aktivitu jako transformační proces (C h e c k l a n d , 1981). S m y t h a C h e c k l a n d (1986) navrhli definici každé aktivity pomocí šesti základní charakteristik – prvků. Tyto prvky se dají použít i ve znalostních tocích, pokud otázky CO-KDO-JAK rozlišíme zvlášť pro tvůrce rozhodnutí a zvlášť pro výkonné složky. Dokumenty z nejrůznějších oblastí se dají strukturovat podle znalostních toků a obsahují-li jasně vymezené části, které dávají odpovědi na základní otázky, dají se tyto části i celé dokumenty automatizovaně vyhledávat pro případ potřeby rychlé odezvy.

strukturované a nestrukturované dokumenty; analýza dokumentů; rychlá reakce; tok znalostí; klíčová slova

Contact Address:

Ing. Ludmila D ö m e o v á , Ph.D., Česká zemědělská univerzita v Praze, Fakulta provozně ekonomická, katedra operační a systémové analýzy, Kamýcká 129, 165 21 Praha 6-Suchdol, Česká republika, e-mail: domeova@pef.czu.cz