Knowledge Discovery and Expertise Location in Unstructured and Semi-structured Data in the Knowledge Management Context

Matej Jambrich
Prague University of Economics, 2005
jambrich@vse.cz

Abstract

The important technological aspect of knowledge management is its support for companies and institutions in their pursuit for structuring, filtrating, contextualization of document bases and additional knowledge artifacts and learning from them. Today, when significant amount of corporate knowledge resides in unstructured heterogeneous data sources (explicit knowledge) and in the minds of individuals (tacit and implicit knowledge), the importance of tools capable of unifying, structuring and contextualizing these isolated “islands”, as well as discovering new connections and structures between them, emerges.

Although the knowledge management and intellectual capital management are heavily discussed topics nowadays, practical functional solutions for enterprise are not very common yet. The knowledge maintenance requires preserving its context. And that is best expressed in the natural language. Therefore most of the contemporary knowledge management systems are built on unstructured or weakly structured documents. Aside from context, the knowledge management is unthinkable without people and their explicitly inexpressible knowledge (the tacit knowledge). The pragmatical conception of knowledge management system must therefore provide the connection between “those who know and those who needs to know”.

The doctoral thesis presents the concept of systems designed for knowledge discovery and expertise location in unstructured and semi-structured data in context of knowledge management (shortly TAEL - Text Analysis / Expertise Location). In the first part of the thesis the concepts of explicit, tacit and implicit knowledge are critically discussed. These are confronted with opinions derived from the state of the art knowledge management literature. Author then proposes his own categorization of software tools supporting knowledge management. The TAEL systems emerge from this categorization.

The main part of the thesis discusses the TAEL systems, the motivators behind these systems and the description of the three main pillars of the TAEL systems - corporate search, expertise location and text analysis. The description of main- and sub-processes, techniques and interfaces used by existing tools for knowledge discovery and expertise location is detailed, after which the general TAEL systems architecture follows. The focus in this part is given to cluster analysis, multidimensional taxonomies, the creation of expert profiles and knowledge maps, as well as to the importance of metadata for TAEL systems. The visualization possibilities and social network analysis are discussed as well.

The second half of the thesis is dedicated to the practical applications of the TAEL systems and the typical structure of the TAEL project. Author describes the interconnections between TAEL and HR systems, TAEL and Business Intelligence and the use of TAEL systems for business processes and forensics support. Additional subchapter talks about the importance of TAEL systems for intellectual capital spreading and consequence with the KBPR (Knowledge Based Process Reengineering) methodology. The rest of the thesis presents the methodology for the TAEL systems evaluation. More than 150 criteria divided into 8 subgroups, as well as
the methodological framework for the selection and criteria importance determination is defined.

The uniqueness of this thesis is in the detailed coverage given to the questions of supporting the knowledge discovery and expertise location. Author believes, that TAEL systems represent the important qualitative shift towards the better use of software tools for knowledge management support.

**Thesis overview**

Outlined on the Figure 1 is the structure of the thesis.

![Figure 1 - Thesis structure](image)

The structure of the thesis stems from the general principles and tends toward the practical implementation of the thesis results. Respective chapters could be divided into the three groups:
- the analysis of the available sources and the analysis of the status quo of the subject area (the inputs),
- detailed analysis of the existing tools for knowledge discovery and expertise location in text, isolation of basic components and building blocks, as well as main and sub processes and their synthesis into the resulting TAEL architecture (the core),
- the common architecture and the detailed description of the TAEL systems, the selection methodology and the proposal of the TAEL practical application (the outputs).

**The outputs**

Due to the limited extent of this abstract, only the outputs in its brief overview are being presented here.

**Architecture**

TAEL systems architecture is defined in two perspectives:

a) process perspective – the main processes together with sub-processes are detailed. The following main processes had been identified:
   a. content extraction,
   b. text preprocessing,
   c. the advanced text processing,
   d. cluster analysis and the knowledge map creation.

b) IT perspective – architecture of the main components and interfaces. Following four layers had been identified:
   a. data connectors,
   b. “backend” services (as defined in the process section),
   c. security layer,
   d. “frontend” services (knowledge portal).

**Practical application**

The proposed TAEL systems architecture is meant to be used as a common conceptual basis for implementation of concrete knowledge management systems. Similarly, the author’s insights from the chapter “TAEL systems application” might serve as a suggestion for practical TAEL systems use. Covered are:

- TAEL and HR integration,
- TAEL and knowledge based process reengineering (KBPR),
- TAEL and forensics (briefly).

Special part of the thesis is dedicated to the TAEL tools selection methodology. More than 150 criteria listed here are gathered into eight groups for better manipulation and selection. These are:

- General, platform, scalability, support,
- Data harvesting,
- Indexing,
- Text analysis,
- Expertise location,
- Search features,
- Knowledge portal and the search quality improvement,
- Operation, service and other.
Author suggests also the multi-criteria evaluation of the variants.

**Conclusion**

The author’s contribution to the subject area could be divided into the contribution for the theory:

- the thesis delivers new and coherent categorization of the tools supporting the knowledge management (“the KM tools hexagon”),
- presents the concept of the TAEL systems as a qualitatively new category and its importance for the knowledge management area,
- the systematical description of the TAEL architecture might be used as a basis for practical knowledge management systems realizations,
- the thesis outlines the areas of TAEL integration with other systems/subdomains,
- and the thesis critically examines the concepts of tacit, implicit and explicit knowledge,

for the praxis:

- the study of the TAEL systems can be used as a guideline for companies and institutions planning to use the knowledge and expertise hidden in and behind their textual bases,
- practical integration of TAEL and other domains,
- description of the typical TAEL systems implementation as well as proposed criteria for measuring the TAEL systems contribution (which could be generalized for other knowledge management implementations),

and for combined areas of theory and praxis:

- the TAEL tools selection methodology suggesting more than 150 criteria generalized into 8 groups.