

Framework for Analyzing and Clustering Short Message Database of Ideas

Mari-Sanna Paukkeri

Helsinki University of Technology, Finland
mari-sanna.paukkeri@tkk.fi

Tanja Kotro

National Consumer Research Centre, Finland
tanja.kotro@ncrc.fi

Abstract: We introduce a framework for a new idea tool NOTE, which gathers, fosters and manages innovative ideas. NOTE supports the development of organizational memory and is connected to the practices of organizational innovativeness. The tool utilizes text mining methods in idea processing, management and visualization and is thus a new approach in idea management software. The tool is under development.

Key Words: idea tool, innovation, text mining, practice, organizational memory

Category: H.2.8, H.3.3, H.4, H.5.2, H.5.3

1 Introduction

Business organizations collect a lot of knowledge, including huge databases about products, customers, competitors, markets, etc., gathered throughout the years. Organizations have also personal knowledge on employee's personal work space, such as text and data files, books, and e-mails, and as an enormous source of information, employee's wide tacit knowledge about their own field and tasks [Nonaka and Takeuchi, 1995], but also hobbyist knowing about many other relevant matters for the organization, outside their own responsibilities [Kotro, 2005]. New ideas and innovations are fundamental in product development of business organizations. Innovations are often based on creativity, but they also arise from the knowledge in an organization. Business organizations have a lot of tacit knowledge, but currently the whole variety of it cannot be used in the innovation process. Furthermore, part of the organizational memory about old ideas disappears every time someone leaves the company.

Innovations are often facilitated with innovation tools but present tools are not good enough to support all the innovativeness and creativity organizations have. One of the reasons is that knowledge is much more diverse than what innovation tools support. There is a need for tools that facilitate communality and common organizational memory evolution in an organization. [Kotro and Paukkeri, 2009]

The innovation process in business organizations can be divided into a number of different stages, including e.g. scenario and portfolio management, business intelligence, idea collection and generation, idea management and evaluation, and project planning [Kohn and Hüsigg, 2006]. In this article, we focus on idea collection, generation, and management.

New ideas occur in the practice of everyday life in organizations and mostly outside the organizational boundaries [Kotro, 2005]. To be able to bring back the ideas occurred outside the working place people tend to write down their ideas and inventions in notebooks, post-it notes and other pieces of paper, or store them to the depths of the file systems of their laptops and mobile phones. Later, it is often impossible to find the note that is needed, or create an overview of all the ideas, not to mention the numerous lost pieces of paper. If this kind of information was collected to an easy-usage electronic notebook it would give a new way to process tacit knowledge in organizations.

Product development team members play with ideas and come up with new ones. An interesting characteristic in revealing ideas to the colleagues is whether the identity of the author is somehow attached to the idea or not. In many situations people tend to be more willing to enter their ideas to an idea collection if they get the praise of having their own name behind the idea. On the other hand, sometimes people prefer to filter their wild ideas if they have to reveal them with their own name. They might feel that they are not experts of the area and do not know all the consequences, and thus the idea might be totally foolish. Also organizational structure might prevent people from mentioning something that is contradictory with other people at a higher position in the hierarchy.

To really support creativity and innovations in organizations, there should exist software tools that collect and remember ideas, manage them, and are easy to use. Such a tool should provide a simple way to enter notes and easy access to the variety of different ideas that are produced by anyone in the organization.

2 Note Tool

NOTE innovation tool helps with the problems of vanishing organizational memory, disappearing or badly accessible notes, and wide range of ideas that are hard to be organized. NOTE is a shared electronic noteboard where the employees of a company write down their observations, ideas and questions. The ideas can be entered whenever they spring into mind, through a mobile phone or a web browser. The underlying data processing system processes the notes and links similar or related ideas together. While contemplating the NOTE database, the arisen idea clusters are visualized in an intuitive way. NOTE may also link external information to the ideas from the databases of the organization or from public databases.

2.1 Use Case

We give an illustrative working example of the NOTE tool from the publishing sector.

A project manager of a company enters a note 'Three of my neighbours have recently bought Kindle' on NOTE. The day after, a product developer adds notes 'People still buy books if they want to read' and 'iPod is used for reading but not for reading entire books'. The NOTE engine links all the related notes to these, generating clusters of books and iPods. The result is visualized as a cloud of notes. The product developer is able to see other related ideas and thus enrich one's own view of the state of the publishing sector (see [Fig. 1]).

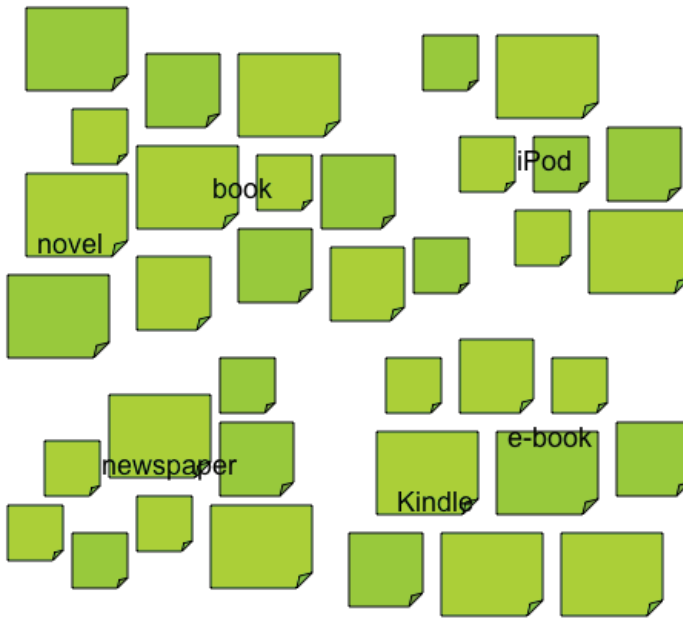


Figure 1: Clustering of notes related to books.

In the afternoon, there is a meeting discussing a possible upgrade from traditional books to e-books. All the ideas relating to e-books in NOTE public database are visualized on the screen. Although lots of notes about electronic books have been added recently, NOTE finds also ten years old notes in the database that suggest paper free offices within a year. That makes people think that the e-book boom might be just momentary hyping as the disappearance of paper was earlier, and they decide not to go out of traditional books.

On the way from a meeting to another, a salesperson writes down the ideas occurred when visiting a customer and submits them to NOTE by SMS. The salesperson checks also the personal NOTE space containing notes about customer companies and their needs. The salesperson creates some additional links between the notes about two companies with similar interests.

2.2 Note Engine

The uniqueness of the NOTE tool is the use of an autonomous engine that processes all the ideas entered to the system. When the number of notes and ideas in the idea collection increases a lot, it starts to be impractical to read through all of them. By using advanced text mining methods, the NOTE engine is able to process a huge amount of data and find semantic connections between the notes. Besides providing the customary lists of ideas, sorted by date, author and category, NOTE provides the user with similar or related ideas that were stored earlier to the idea database of the company. The NOTE engine groups and clusters ideas according to the semantic similarity of the texts. NOTE remembers older ideas and picks up the relevant ones from the large idea collection. The tool visualizes ideas in a way that a clear view of the entirety of the needs, innovations under process and targets of development in the company can be established.

NOTE adds a date stamp and either user name or anonymous tag to each note entered to the system. Thus temporal changes can be tracked and statistics given about the emergence of innovations. Users may create and delete links and see clusters according to their own preferences. In addition to the public workspace, each user has a private workspace where they can develop and collect their individual ideas. In a later version, NOTE will link customer data, visualizations and other relevant external information to the notes.

2.3 Text Mining Methods

The NOTE engine uses statistical text mining methods for processing notes. Short notes, about 1–20 words each, may be written in any language for which there is also textual background material available. The background material is used as a sample of 'general language' for the clustering methods. In the first version of the NOTE engine, first and second order term-document matrices are used together with *tf.idf* score [Salton and Buckley, 1988] to create a semantic vector space of the idea texts. The clustering is conducted in the vector space by *k-means* algorithm [MacQueen, 1967] with cosine distance measure.

Language is redundant with synonyms and ambiguity and thus semantically similar short notes share the same vocabulary very rarely. Therefore, the second version of the NOTE engine will use a bias corpus: the bias corpus is a collection

of text documents in the used language and on the area of the business (for example, from the company's own databases or publicly available data). The bias corpus will be used for expanding the short notes, to be able to calculate the similarities between notes more accurately.

The NOTE text mining methods are statistical, as the opposite to rule-based systems with fixed vocabulary and restricted sentence structure, and they follow the unsupervised machine learning paradigm. Due to the statistical approach, NOTE is able to process new, not-predefined words and phrases by simply running the unsupervised clustering again. NOTE text mining methods are also language independent, that is, basically the system needs only reference corpus in the used language [Paukkeri et al., 2008] and is thus easily modified to a new language. Anyway, to obtain more smooth performance, some small, easily replaceable language-specific components must be added. Text mining algorithms will be discussed in more detail in a forthcoming publication.

2.4 Visualization

The clustering of NOTE may be local, linking just the most relevant notes to the idea at hand, or global, organizing all the notes in the databases into clusters. The unsupervised clustering is visualized according to the preferences of the user:

- In **General overview**, all the publicly available notes in the database are clustered and arranged according to the semantic similarity. The general view helps in tracking the emphasis of ideas in the company and the changes over time.
- **Notes on one's personal space** can be clustered separately or combined to the public notes.
- **Notes related to certain note or topic** can be linked to build a local cluster.
- **Classification** according to a predefined list of product names or keywords, or automatically extracted keyphrases from an external set of documents [Paukkeri et al., 2008].

The user may facilitate the cluster to have the emphasis on:

- **Recent notes**; notes with a recent time stamp get more weight.
- Notes by certain **author**.
- According to **votes** given to notes.

The visualization is based on the resulting clusters of notes and the cosine distance between notes in the term-document vector space. The growing database does not cause trouble to the visualization, since links to the most relevant notes only are visualized in the local clusters. An important part of the visualization is to select appropriate labels for the clusters and sub-clusters in the global clustering [Lagus and Kaski, 1999].

The framework of NOTE tool is visualized in [Fig. 2]. The system contains a large database of notes, a bias corpus and external information. The clustering engine performs local and global clustering for the notes and gives the output to the visualization engine. The users may use NOTE freely by mobile phone or through Internet.

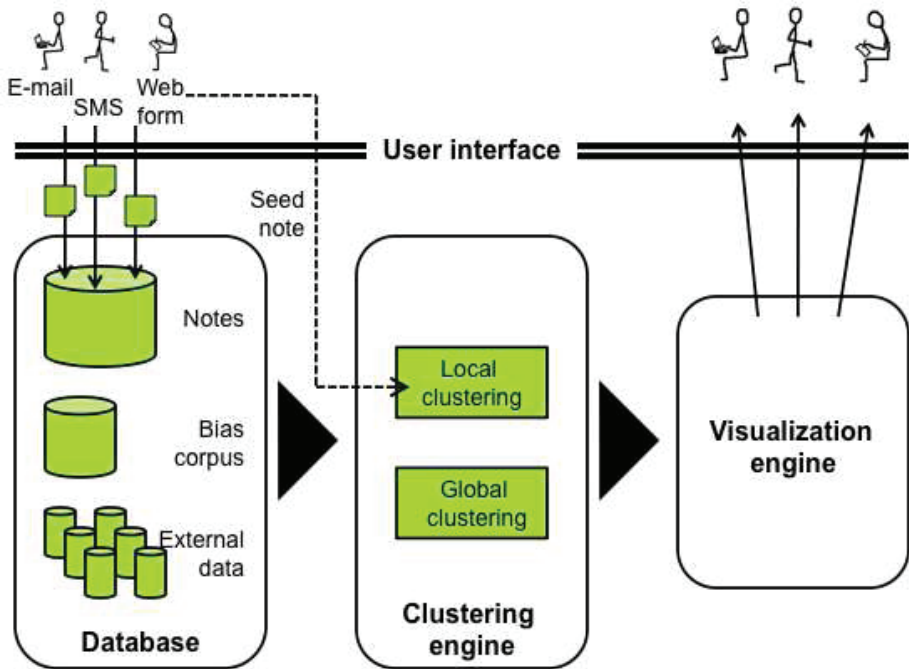


Figure 2: The NOTE idea tool framework.

3 Related Work

This work is multidisciplinary work and has connections to several fields, ranging from organizational theory research through knowledge management to text

mining. Organizational theory, and especially innovation as a decentralized activity in organizations, forms an important background for the key ideas of the NOTE tool [Kotro, 2005, Kotro, 2007, Kotro and Paukkeri, 2009]. Text mining discussion will come as future work.

There are lots of idea generation and management tools available on the market. Many of the tools are based on mind maps or trees and their visualization. The user is able to write their own mapping and scan through the previous visualizations. Usually there are some functionalities to process the created map structures. The nodes of the maps contain usually few words. Some examples are *Decision Explorer* (Banxia Software Ltd.)¹, *Idea processor* (Axon Research)², and *MindView 3* (MatchWare)³. These systems are easy and effective to use but they are just electronic versions of the pencil-and-paper mind maps and do not involve any processing of the texts or comparisons between maps.

Goldfire (Invention Machine)⁴ is innovation software that supports several innovation tasks and tools. It uses semantic indexing technology to extract relevant concepts from internal and external databases, like corporate documents, content repositories and technical websites. Concepts are collected to knowledge bases to allow *Goldfire* to generate dynamic taxonomies of query results and create summaries of documents. *Goldfire* processes English, French, German and Japanese text and also supports translations between these languages. The system analyzes documents with computational linguistics methods but does not directly utilize those methods in innovation refining.

Some approaches for clustering and visualizing document collections have been proposed. *WEBSOM* [Honkela et al., 2004, Kohonen et al., 2000] organizes vast document collections to a two-dimensional plane using statistical representations of the vocabulary used in the documents. The clustering method is *Self-Organizing Map (SOM)*, an artificial neural network method that uses unsupervised learning paradigm. *WEBSOM* is used to analyze e.g. patent abstracts. *eClassifier* [Cody et al., 2002] categorizes automatically large collections of text documents. It has many different visualization techniques and also extraction techniques to provide concept summarizations for each category. *SWAPit* analyzes and visualizes document collections using ontologies as the basis of clustering [Seeling and Becks, 2004]. It uses a multiple views paradigm to enable user to simultaneously explore documents on content level, conceptual level and based on associations to structured data records. The reviewed visualization methods use article abstracts or whole articles as their data. For shorter texts, such as notes, different methods are needed. Methods to analyze shorter contexts are discussed in [Pedersen, 2008].

¹ <http://www.banxia.com/dexplore/index.html>

² <http://web.singnet.com.sg/~axon2000/>

³ <http://www.matchware.com/en/products/mindview/default.htm>

⁴ <http://www.invention-machine.com/ProductsServices.aspx?id=50>

4 Conclusions and Future Work

We introduced an idea of a tool that facilitates innovativeness in a business organization. The NOTE tool is used to create and update the collective memory of a company. The tool combines statistical text mining methods with the daily practices of an organization in a novel way.

NOTE combines different sources of information and brings unofficial and practical information into a part of product development process and decision-making. It emphasizes creative brainwork, supports virtual communities and face-to-face teamwork, and helps product managers to innovate. NOTE lowers the hierarchy in an organization by collecting innovations from every level of the hierarchy, according to the value of the idea, not the status of the employee.

A demonstration version of the system has been implemented with most of the functionalities: web page and e-mail submission of notes, the first version of clustering and most of the visualization features. The future of the NOTE framework includes the development of a working application with the second version of the NOTE engine that utilizes bias corpora. Furthermore, qualitative research of the usage of NOTE in business organizations will be conducted. Further technical steps after the second version of the NOTE engine are adding some extra functionality, e.g. machine translation that would be useful in international organizations.

Acknowledgements

This work has been supported by the Finnish Funding Agency for Technology and Innovation (*TEKES*), within the *KULTA2* project.

References

- [Cody et al., 2002] Cody, W. F., Kreulen, J. T., Krishna, V., and Spangler, W. S. (2002). The integration of business intelligence and knowledge management. *IBM Systems Journal*, 41(4):697–713.
- [Honkela et al., 2004] Honkela, T., Nordfors, R., and Tuuli, R. (2004). Document maps for competence management. In *Proceedings of the Symposium on Professional Practice in AI*, pages 31–39. IFIP.
- [Kohn and Hüsigg, 2006] Kohn, S. and Hüsigg, S. (2006). Potential benefits, current supply, utilization and barriers to adoption: An exploratory study on german smes and innovation software. *Technovation*, 26(9):988–998.
- [Kohonen et al., 2000] Kohonen, T., Kaski, S., Lagus, K., Salojärvi, J., Honkela, J., Paatero, V., and Saarela, A. (2000). Self organization of a massive document collection. *IEEE transactions on neural networks*, 11(3):574–585.
- [Kotro, 2005] Kotro, T. (2005). *Hobbyist Knowing in Product Development: Desirable Objects and Passion for Sports in Suunto Corporation*. PhD thesis, University of Art and Design, Helsinki, Finland.
- [Kotro, 2007] Kotro, T. (2007). User Orientation Through Experience: A Study of Hobbyist Knowing in Product Development. *Human Technology*, 3(2):154–166.

- [Kotro and Paukkeri, 2009] Kotro, T. and Paukkeri, M.-S. (2009). Designing a software tool for organizational creativity - the challenge of the practice. Forthcoming.
- [Lagus and Kaski, 1999] Lagus, K. and Kaski, S. (1999). Keyword selection method for characterizing text document maps. In *Proceedings of ICANN99, Ninth International Conference on Artificial Neural Networks Artificial Neural Networks*, volume 1, pages 371–376.
- [MacQueen, 1967] MacQueen, J. (1967). Some methods for classification and analysis of multivariate data. In *5th berkeley symposium*, volume 1, pages 281–297.
- [Nonaka and Takeuchi, 1995] Nonaka, I. and Takeuchi, H. (1995). *The knowledge-creating company: How Japanese companies create the dynamics of innovation*. Oxford University Press, USA.
- [Paukkeri et al., 2008] Paukkeri, M.-S., Nieminen, I. T., Pöllä, M., and Honkela, T. (2008). A language-independent approach to keyphrase extraction and evaluation. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 83–86, Manchester, UK.
- [Pedersen, 2008] Pedersen, T. (2008). Computational approaches to measuring the similarity of short contexts: A review of applications and methods. *South Asian Language Review*, in print.
- [Salton and Buckley, 1988] Salton, G. and Buckley, C. (1988). Term weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523.
- [Seeling and Becks, 2004] Seeling, C. and Becks, A. (2004). Analysing associations of textual and relational data with a multiple views system. In *CMV '04: Proceedings of the Second International Conference on Coordinated & Multiple Views in Exploratory Visualization (CMV'04)*, pages 61–70. IEEE Computer Society.