Agarwal, N.K. & Poo, D.C.C. (2006). Making sense of an electronic document: Visualization strategies for concept presentation. In *Proceedings of the 10th IEEE International Enterprise Distributed Object Computing Conference Workshops, Hong Kong, Oct 16-20* (pp. 56-59). Los Alamitos: IEEE Computer Society. https://doi.org/10.1109/EDOCW.2006.45

Making sense of an Electronic Document – Visualization Strategies for Concept Presentation

Naresh Kumar Agarwal and Danny C. C. Poo
School of Computing, National University of Singapore
3 Science Drive 2, Singapore 117543
{naresh, dpoo}@comp.nus.edu.sg

Abstract

Electronic documents are increasingly becoming the norm today. Electronic document management systems solve many of the storage and retrieval problems inherent in paper filing systems while reducing business costs. However, with conversion of paper documents to electronic form, and an increasing amount of data being produced in electronic form, an employee today is hard-pressed in time simply trying to make sense of all he is required to read on his monitor. In order to help manage information overload and to help 'connect the dots' among various pieces of information, it is important that mechanisms and tools be developed for sieving the important bits of information, collecting and assimilating them in a useful manner and presenting them for easy digestion. In this paper, we discuss 4 visualization strategies – 1) concept/mind maps 2) taxonomies/facets/ontologies 3) questions and answers (Q&A) and 4) summarization that should be adopted to help present the concepts in an electronic document for easy comprehension. Such strategies will go a long way in cutting down the reading time of the employee, leading to productivity increases and cost savings for the enterprise.

1. Introduction

Electronic documents are increasingly becoming the norm today. From documents to research papers to books to emails to reports, we have enough to keep our eyes glued to computer screens. A study conducted at the University of California, Berkeley [1] concluded that print, film, magnetic and optical storage media produced about 5 exabytes (equivalent to all words ever spoken by human beings) of new information in the year 2002 alone. With a world population of 6.3 billion, almost 800 MB of recorded information is produced per person each year. In addition, information flows through electronic channels (telephone, radio, TV and the Internet) contained almost 18 exabytes of new information in 2002, 3.5 times more than is recorded in storage media. Also, worldwide production of new information recorded on magnetic storage media has grown 87% since 1999. Compared to the growth of the World Wide Web, 'development of the human brain has been tardy: it has grown only linearly from 400 to 1400 cubic centimeters in the last 3.5 million years' [2].

Electronic Document Management (EDM) is 'the application of technology to save paper, speed up communications, and increase the productivity of business processes.' With the application of technology to documents and document processing, EDM promises major productivity and performance increases [3] in the workplace. Such systems solve many of the storage and retrieval problems inherent in paper filing system while reducing business costs.

However, with conversion of paper documents to electronic form, and an increasing amount of data being produced in electronic form, an employee in an enterprise today is hard-pressed in time simply trying to make sense of all he is required to read on his monitor. There may be more text in electronic form than ever before, but much of it is ignored. No human can read, understand, and synthesize megabytes of text on an everyday basis. Missed information, and lost opportunities, has spurred researchers to explore various information management strategies to establish order in the text wilderness [4]. In order to help manage this information overload and to help *connect the dots* among various pieces of information, it is important that mechanisms and tools be developed for sieving the important bits of information, collecting and assimilating them in a useful manner and presenting them for easy digestion. This would be of immense benefit to researchers as well as practitioners. Filtering techniques, search engines, taxonomies and

ontologies are all efforts in this direction but the challenge of information overload remains.

In this paper, we discuss visualization strategies that should be adopted in EDM systems to help present the concepts in an electronic document for easy comprehension. Specifically, we discuss:

1. Concept and mind maps,
2. Taxonomies/facets/ontologies,
3. Questions and answers (Q&A) and
4. Summarization/Abstract

Such strategies will go a long way in cutting down the reading time of the employee, leading to productivity increases and enterprise cost savings. In Section 2, we look at the four visualization strategies. In Section 3, we briefly discuss the practical utility of the strategies, as well as the issues to keep in mind when implementing these strategies. In Section 4, we see an example of the strategies applied to a specific document. We conclude the paper in Section 5. Let us now look at the visualization strategies.

2. Visualization Strategies for Concept Presentation

The goal of information visualization is the unveiling of the underlying structure of large or abstract data sets using visual representations that utilize the powerful processing capabilities of human visual perceptual system [5]. The aim here is to help cut down on the reading time of an individual and to help him assimilate information easily with the help of presentation outputs such as concept maps, taxonomies and Q&A views. Once the key ideas are understood, the reader could then go ahead and read the full paper/document depending on need and importance (with an increased understanding than before). These views would serve the utility of notes a student makes from a text, for quick review before exams. It would also be akin to seeing a PowerPoint presentation of a particular topic before reading the full paper. Let us now look at the four visualization strategies.

2.1. Concept and Mind Maps

Mind map or concept map has been used in various kinds of situations. It was first developed in the late 1960s by Tony Buzan as a way of helping students make notes that use only keywords and images. Lately, software has been developed to help people to visualize information and their ideas e.g. SmartDraw (www.smartdraw.com), visual-mind (www.visual-mind.com), etc. However, all these tools are pure graphic drawing tools. Software packages like CmapTools (http://cmap.ihmc.us), which facilitate the online manipulation of concept maps, have extended their use and applicability to knowledge sharing, organization and browsing [6]. Canas *et al.* [7] present an algorithm for linking concepts into meaningful propositions creating a coherent structure that reflects the learner's understanding of a domain. An effective EDM system must be able to automatically generate a concept map of the key ideas of a document.

2.2. Taxonomies/ Facets/ Ontologies

Taxonomies have been used for quite some time to hierarchically represent and categorize information. A taxonomy provides a subject-based classification that arranges the terms in a controlled vocabulary into a hierarchy. Humans can rapidly navigate taxonomies to find high concentrations of topic-specific, related information [8][9][10]. *Faceted classifications* work by identifying a number of facets into which the terms are divided e.g. classifying by color, geography, subject, etc. Facets can be thought of as different axes along which concepts in a document can be classified, and each facet contains a number of terms. This would then describe the document from many different perspectives [10]. Taxonomy is a type of facet in which the headings are arranged into a hierarchy.

While taxonomies present a broader/narrower relationship in building a hierarchy, *ontologies* help specify broader, networked relationships. Taxonomies, facets and ontologies can serve as useful ways to classify the key concepts in a document for easy navigation by the busy professional.

2.3. Questions and Answers (Q&A)

The Question and Answer (Q&A) presentation view shows the possible questions that can be formed from the text of the document and their answers. While concept maps and taxonomies are actually similar, as they are different ways of representing the key concepts of a document, the Q&A is quite different. The idea here is that given one or more sentences in a document, the EDM system should automatically come up with a question that is answered by the sentence(s) in the document. A person reading could pick and choose the questions of interest and sieve out the unimportant ones.

Question Answering (QA) research attempts to deal with a wide range of question types including, fact, list, definition, How, Why, hypothetical, semantically-

constrained, and cross-lingual questions. QA systems such as BASEBALL, LUNOR, SHRDLU and ELIZA have been developed since the 1960s. Other systems such as the Unix Consultant (UC) and LILOG were developed in the 1970s and 1980s. Since the late 1990s, the annual Text Retrieval Conference (TREC) has included a question-answering track, where participating systems are expected to answer questions on any topic by searching a corpus of text that varies from year to year. Currently, there is an increasing interest in the integration of question answering with web search. E.g. Ask.com is one such system. Google and Microsoft are also venturing into QA. (Wikipedia[1])

A good EDM system should incorporate not just answering of questions based on the document, but formulation of questions based on textual data of the document.

2.4. Summarization/Abstract

Abstract is a short summary of any given text/document. A summary implies extracting some sentences or keywords from a document. However, a computer cannot link words, sentences together to make meaningful content after extraction. A lot of prior research has been done in the area of summarization. E.g. Mani and Bloedorn [11] propose a summarizing method based on a graphic representation of related documents. This method requires deep analysis of the original text. Salton *et al.* [12] suggest passage/paragraph extraction from a document based on intra-document links between paragraphs. A good summarizing capability should be an important attribute of an EDM system.

3. Practical Utility of Strategies and Implementation Issues

Table 1. Utility of Visualization Strategies

Visualization Strategy	Utility
Concept / Mind Map	Will help the reader get an overall feel and to help make sense
Taxonomy / Tree View	Will give a top-down view of the document with the key points bulleted
Q&A View	Ideal for employees who want to see documents from a question-centric angle. They can mark/extract and concentrate on important questions.
Abstract / Summary	Will help the reader get a quick overview of what lies ahead

[1] http://en.wikipedia.org/wiki/Question_answering

Table 1 lists the utilities of the four visualization strategies. E.g. instead of jumping into reading a 40-page report, an employee in a firm will be able to read a one-paragraph summary of the report, see a conceptual view of the entire document in the form of mind/concept maps, see a hierarchical taxonomy view of the document and view the document in the form of questions and answers (and focus on those questions he is interested in). If still needed, the employee can then read the relevant parts of the report after getting a quick overview.

While implementing these strategies, the EDM system must allow the user to specify the granularity level of detail he wants in the output generated of the document. The user must also have the flexibility of choosing the regions of interest in the document. He must be able to refine the concept map or taxonomy view and to zero in and choose the root of the concept map for greater accuracy. He should be allowed to click on the nodes of the concept map to expand/collapse the details of the node. This would give him more clarity on the document. Such a system must also allow saving of presentation preferences for future use. The EDM system must be able to capture presentation preferences of the user by noting his relevance feedback. Ideally, the user must also be able to navigate between the presentation output and the original location in the input document. E.g. while viewing the mind/concept map, a user could right click on any node and go to the location of this concept node in the original document. The system must also have the provision for specifying classes of documents where a certain strategy may be unsuitable e.g. documents that already have an abstract will not require a system-produced abstract.

4. An Example – Application of Strategies to a specific document

Given a sample document on 'Multimedia Messaging Service (MMS)' (accessible at http://en.wikipedia.org/wiki/Multimedia_Messaging_Service), a clickable concept map of the document would look like the one shown in Figure 1. Figure 2 shows the concept map with one node open on clicking.

Figure 1. Clickable Concept Map of Document

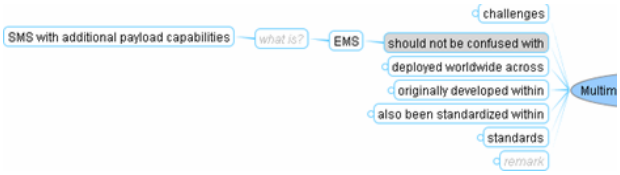

Figure 2. One Node Opened

Similarly, any number of nodes could be opened or closed as per user convenience. The user should also be able to double click on a particular node and go to the specific area of the electronic document. Figure 3 shows a Taxonomy View of the same document.

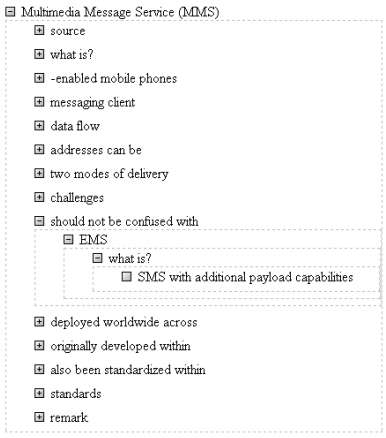

Figure 3. Taxonomy View

A sample of the Q&A view could look like the one shown in Figure 4. A summary of the same document could also be generated.

Q) What is MMS?
MMS is the logical evolution of the *Short Message Service* SMS, a text-only messaging system for mobile networks

Q) What are the two types of delivery modes in Multimedia Messaging Services (MMS)?
Immediate delivery or deferred delivery

Figure 4. Sample of Q&A View of Document

5. Conclusion

We have described four visualization strategies that should be adopted to help present the concepts in an electronic document for easy comprehension. The application of these strategies to Electronic Document Management systems should help cut down on the reading time of an individual and help him assimilate information easily. This will lead to productivity and performance increases in the work place, and increase the efficacy of electronic document management.

References

[1] K. Swearingen, P. Lyman, H.R. Varian, P. Charles, N. Good, L.L. Jordan, and J. Pal, "How much information? 2003", *School of Information Management and Systems*, University of California, Berkeley, 2003, Retrieved 7 Jul 2006 from http://www.sims.berkeley.edu/research/projects/how-much-info-2003/

[2] S. Chakrabarti, M. Berg, and B. Dom, "Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery", *Computer Networks*, Vol. 31, 1999, p.11-16, Retrieved 7 Jul 2006 from http://www.cse.iitb.ac.in/~soumen/doc/www1999f/pdf/www1999f.pdf

[3] R.H. Sprague, Jr., "Electronic Document Management: Challenges and Opportunities for Information Systems Managers", *MIS Quarterly*, Vol. 19 Issue 1, March 1995, pp.29-49.

[4] J. Cowie, J., and W. Lehnert., "Information Extraction", *Communications of the ACM*, Vol. 39 Issue 1, 1996, pp.80-91.

[5] M. Hearst, "SIMS 247: Information Visualization and Presentation", SIMS, Berkeley, 2004, Retrieved 7 Jul 2006 from http://www.sims.berkeley.edu/academics/courses/is247/s04/

[6] M. Carvalho, R. Hewett, and A.J. Canas, "Enhancing Web Searches from Concept Map-based Knowledge Models", SCI Conference, July 22-25, 2001, Orlando, Florida.

[7] A.J. Canas, M. Carvalho, and M. Arguedas, "Mining the Web to Suggest Concepts during Concept Mapping: Preliminary Results", XIII Simposio Brasileiro de Informatica na Educacao, SBIE, UNISINOS, 2002.

[8] P. Lederman, "Implementing a Taxonomy Solution", *AIIM E-Doc*, Vol. 19 Issue 2, Mar/Apr 2005, pp.25-26.

[9] A. Papadopoullos, "Answering the Right Questions about Search", *EContent Leadership Series - Strategies for...Search, Taxonomy & Classification*, Supplement to July/August 2004 EContent and Information Today, pp. S6-S7, Retrieved 6 Jul 2006 from http://www.procom-strasser.com/docs/Convera_Right_Questions.pdf

[10] L.M. Garshol, "Metadata? Thesauri? Taxonomies? Topic Maps! Making sense of it all", *Ontopia*, 2004, Retrieved 6 Jul 2006 from http://www.ontopia.net/topicmaps/materials/tm-vs-thesauri.html#N828

[11] I. Mani, and E. Bloedorn, "Multi-document summarization by graph search and matching", *Proceedings of the Fourteenth National Conference on Artificial Intelligence (AAAI-97)*, July 27-31, 1997, Providence, Rhode Island, pp.622-628.

[12] G. Salton, A. Singhal, M. Mitra, and C. Buckley, "Automatic text structuring and summarization", *Information Process Management*, Vol. 33 Issue 2, March 1997, pp.193-207.