AI support by Intelligent Documents

A.W.A. (Sander) Spek H. Jaap van den Herik

Institute for Knowledge and Agent Technology (IKAT), Universiteit Maastricht,
Faculty of General Science, P.O. Box 616, 6200 MD Maastricht

Abstract

In this paper, we start with a statement on why adequate artificial-intelligence techniques do not reach the business audience. We aim at improving the link to business applications by presenting a framework of techniques that is useful for better knowledge distribution. Intelligent documents are the backbone of this framework. Our contribution is in combining classification and personalisation techniques to the framework of intelligent documents.

1 Introduction

Innovation relies on two factors: technology supply and user demand. Weggeman [16, p.61] identifies these factors as *technology push (product-out)* and *market pull (market-in)*. In practice, either side may take the initiative. Both sides incrementally add pieces of innovation (e.g., innovative ideas) to the 'innovation market'. Figure 1 displays the relation.

Technology and innovation researcher Smits acknowledges "there can be a great gap between R&D and the application of it" [15, p.27]. Despite all recent technical, rather advanced, innovations in artificial intelligence, the applications of these techniques in practice still falls behind. Companies only rely on proven technologies carried out by big and —financially— stable software houses. Pre-procurement software-evaluation documents in a multinational company support this statement. So far, innovative artificial-intelligence techniques from scientific research are frequently not exposed to a big company audience; certainly not without taking the long route via established software houses.

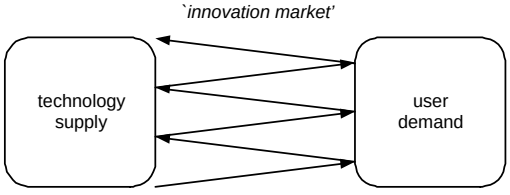

Figure 1: The 'innovation market', in which technology supply and user demand enhance each other to reach higher innovation levels.

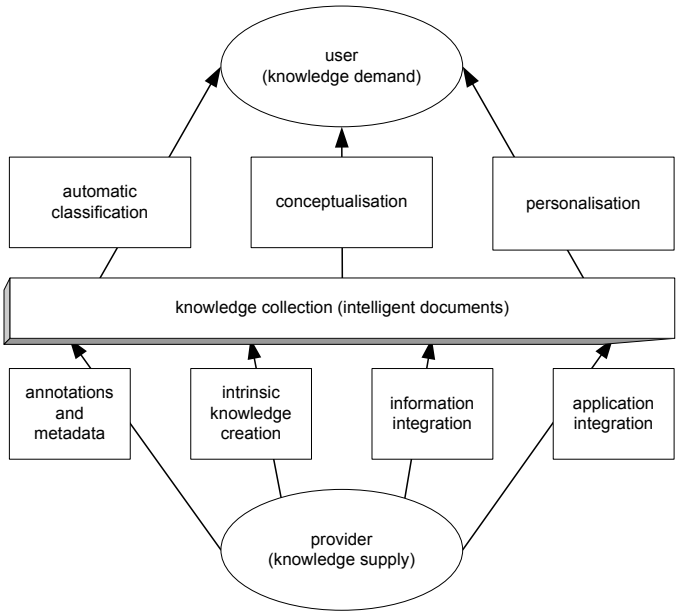

Figure 2: Framework for artificial intelligence support in knowledge sharing via collections

In this article, we will link innovative AI technology to the daily practice of large companies. We start by presenting a framework which contains a survey of AI techniques with a potential in aiding users at finding the required knowledge in large knowledge collections (section 2). Intelligent documents (section 3) are the core of this framework. Subsequently, we will discuss in detail the possibilities of two interesting techniques, and their relation to intelligent documents. The techniques are automatic classification (section 4) and personalisation (section 5). Conclusions are given in section 6.

2 Framework

In [13], we identified seven categories of problems that users may encounter when searching for information in multi-domain collections. In the framework of figure 2, we provide seven AI-oriented solutions for these categories.

The framework has two actors: (1) a provider, and (2) a user; a knowledge collection lies in between. The knowledge provider adds information to the collection, whereas the user has a demand, usually requesting knowledge. In the knowledge collection all the codified knowledge of the system is stored.

The seven techniques help the user to facilitate knowledge retrieval directly or indirectly. Four of them (viz. annotations and metadata, intrinsic knowledge creation, information integration, and application integration) do this by

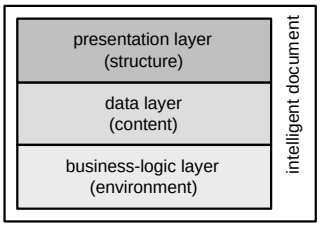

Figure 3: The layers of an intelligent document

modifying the knowledge offered to the collection. For instance, intrinsic knowledge creation applies statistical analysis methods to find relationships amongst documents. These four techniques, in particular annotational data and intrinsic knowledge creation, provide a step towards intelligent document collections. Three techniques (viz. automatic classification, conceptualisation, and personalisation) modify (the representation of) the knowledge when going from retrieval to the user presentation. These techniques are aided by intelligent documents.

In the next section, we discuss the knowledge collection by a recent development: *intelligent documents*. Then automatic classification and personalisation are selected as the techniques to be the most interesting for further study. So we will discuss them in separate sections (4 and 5), along with their relation to intelligent documents.

3 Intelligent documents

We define an intelligent document as a document that contains in-depth knowledge about itself and its environment. It is a higher level of annotational data. An intelligent document consists of knowledge on (1) structure, (2) contents, and (3) context (environment) [1].

Adobe has developed the idea of intelligent documents into their Intelligent Document Platform [10]. It facilitates the use of Adobe's Portable Document Format for intelligent documents. In Adobe's terminology, the intelligent document consists of a presentation layer, a content layer, and a business-logic layer. See figure 3.

The business-logic layer controls the context of the document, e.g., by allowing validation rules, workflow-routing information, and security levels to be added to it. The content-layer allows the capture of knowledge into a standard and facilitates its transportations amongst people and applications. The presentation layer allows the document to contain a rich content presentation. Some examples are lay-out issues, a customized order of information elements, and the inclusion of different kinds of media.

The adoptation of intelligent documents by Adobe makes it available to business users. In the Netherlands, Adobe co-operates with amongst others DSoft, SAP, IBM, and Accenture to create a platform for the technique [10].

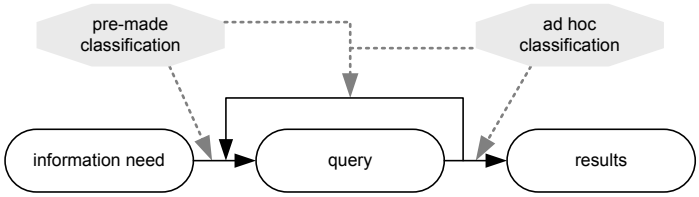

Figure 4: Applying classifications to an information search

4 Automatic classification

Automatic classification refers to techniques of sorting documents in classes. We can define it as the process of dividing a set of information sources into classes (or groups, or clusters), where the sources in one class have certain relevant features in common.

Dumais and Chen [4] distinguish flat and hierarchical classifications. Flat classifications are straightforward divisions. In hierarchical classifications, categories can be ordered in a tree, creating supercategories and subcategories. In both cases, the collection of categories is a taxonomy [5, 9].

Presumabely because of the clear advantages over human classification (see [3]), the first notions (at least to our knowledge) of automatic classification date back to 1979. Van Rijsbergen [14] identified two main areas of application of classifying methods in IR: (1) keyword clustering, and (2) document clustering.

4.1 Use of classification

During an electronic information search process there are several stages where a classification can prove beneficial. We provide three instances (figure 4). The first instance describes a stage between the appearence of an information need and the creation of a query. The user can then select one class (or perhaps multiple classes) to which the search will be limited. The second instance deals with a stage during which a query is reformulated. When a query is suggested to the system, it can turn out that one or more terms are ambigious. The system can then return to the user to check which meaning was meant, by allowing the user to select a class to apply the search on. The third instance describes classification which is beneficial when sorting the retrieved document, to present them to the user in a structured manner.

HTML documents, and their ability to link to other documents, allow another kind of classification, based on these links [6, 11].

4.2 Classification and intelligent documents

Intelligent documents can help classification techniques to deliver more added value to the user. The separation of content from context and structure allows better classification based on the content. The context and structure can separately be

used to discover other features on which the documents can be classified. The essence of this classification supported by intelligent documents is its versatility. Classifications can be made according to coincidental features that make up the environment. Such a classification can, for instance, be used for detection of rules that are obstacles for international commerce.

5 Personalisation

Personalisation provides the user with information based on his own preferences or profile. It assumes that every user is unique, and claims that consequently also information preferences are unique. The term is often confused with customisation (allowing the user to customise, i.e., some interface), but is yet more powerful.

Users generally are not good at formulating adequate search queries. However we believe that despite the poor communication, specific information can be selected in advance, with the aim to provide the users with information tailored to their preferences.

In the next sections, we discuss the methods of collecting profile data (subsection 5.1) and profile types (subsection 5.2), and show the value of intelligent documents (subsection 5.3).

5.1 Collecting profile data

To compose a profile, we need to collect and store a user's preferences. Then we can anticipate on these preferences. A collection of preferences (or data from which these preferences can be derived) is called a *profile*. Below, we briefly describe methods to collect data for such a profile

The most direct way to collect data for profiling is by performing relevance feedback [2]. In [12], an example is provided by showing a real estate agent using relevance feedback to acquire knowledge on the user's wishes. A big disadvantage, that makes relevance feedback almost unusable in some cases, is that continuous user intervention is required to score items. To avoid this problem methods of collecting data indirectly have been developed.

5.2 Profile types

After the data for the profiles has been collected, the system has to interpret them. Three types of profiles, with associated profile analysis, can be distinguished: (1) specifying rule-based profiles, (2) content-based filtering, and (3) collaborative filtering. Also hybrid forms are possible.

5.2.1 Rule-based personalisation

Rule-based personalisation is a method of personalisation where decisions are based on certain rules. This is the most basic option for personalisation. The user, or an expert, can specify a set of rules to score objects. These rules can

be straightforward, like a single keyword, or a balanced complex using multiple keywords and boolean operators.

A drawback of this method is that rules have to be created. It can be done manually, or by automatic extraction. For instance, geographical locations can be derived from IP-addresses. Such rules can consequently be used when filtering or adding certain elements from or to the set of returned information.

5.2.2 Content-based personalisation

Content-based personalisation, or induction, composes a profile out of the user's history of past interests (e.g., [7, 17]). This often involves logging the user's actions.

A major advantage of this method is that users can get insight into the motivation why items are considered interesting for them. They are expected to like document A because they also liked document B and C. Two main drawbacks can be identified. First, the fact that outlying items might cause problems: they will not be recommended, and they are of no help for the recommendation of other documents. Second, content-based personalisation is genre specific. A system that knows a user's preferences in computer-science literature has no clue what kind of fiction novels the user likes.

5.2.3 Collaborative filtering

Collaborative filtering, or deduction, gathers knowledge on general user traits, and models this into a profile [17]. A profile is compared to other profiles by detecting similarities and opposities. It then makes predictions upon these comparisons.

Compared to content-based personalisation collaborative filtering has some potential to cross genre borders. However, it does not give acceptable insight to the user into motivation behind recommendation. A collaborative-filter system biases popular items, since they get recommended more often. It will make them more popular again, and the circle repeat. Outlying users cannot be covered by the system, and also items must be reviewed several times before they can be recommended. Finally, recommender systems face the possible problem of having many items and only a few users [17].

Personalisation, or so-called 'recommender systems', have some successful implementations. For example, many people know them from e-commerce webshops, such as Amazon.com, where products are recommended. Another succesful example is the Grouplens project by the University of Minnesota (i.e., see [8]). A derivative of this project is Movielens[1], a movie-recommender system that is free to use for the Web audience.

5.3 Personalisation and intelligent documents

Obviously, intelligent documents support the personalisation mechanism. An appropriate isolation of content from other elements allows a better analysis of the

[1] `http://movielens.umn.edu`

content and is therefore benficial to a combination with content-based personalisation. Environmental knowledge can also be quite important for personalisation. Business rules, like document workflows, give valuable indications whether documents should be in a person's field of interest. Especially when distributing knowledge via a push-mechanism (versus a pull-mechanism), workflows included in documents are of great aid when combined with role-based personalisation and collaborative filtering.

6 Conclusions and future work

Our framework (figure 2) shows the value of seven AI techniques to everyday knowledge sharing. Intelligent documents are the core of the framework. We have shown the added value of intelligent documents to classification and personalisation.

The growing importance of knowledge sharing as well as intelligent documents asks for continued research. The practical implementation of intelligent documents in relation to the AI techniques will be performed in the near future and the implication on the resulting user benefits will be researched.

References

[1] H. Ahonen, B. Heikkinen, O. Heinonen, J. Jaakkola, P. Kilpeläinen, G. Lindén, and H. Mannila. Intelligent assembly of structured documents. Technical Report C-1996-40, University of Helsinki, Department of Computer Science, 1996.

[2] W. B. Croft, S. Cronen-Townsend, and V. Lavrenko. Relevance feedback and personalization: A language modeling perspective. In *Proceedings of the DELOS-NSF Workshop on Personalization and Recommender Systems in Digital Libraries*, June 2001.

[3] Delphi Group. Taxonomy & content classification: Market milestone report. White Paper, 2002.

[4] S. T. Dumais and H. Chen. Hierarchical classification of web content. In *Proceedings of SIGIR'00*, pages 256–263, August 2000.

[5] A. Fall. *Reasoning with Taxonomies*. PhD thesis, Simon Fraser University, 1996.

[6] G. W. Flake, S. Lawrence, C. L. Giles, and F. Coetzee. Self-organization of the web and identification of communities. *IEEE Computer*, 35(3):66–71, 2002.

[7] H. R. Kim and Ph. K. Chan. Learning implicit user interest hierarchy for context in personalization. In *Proceedings of Intelligent User Interfaces (IUI'03)*, pages 101–108, Miami, Florida, USA, January 2003. ACM.

[8] J. Konstan, B. Miller, D. Maltz, J. Herlocker, L. Gordon, and J. Riedl. Grouplens: Applying collaborative filtering to usenet news. *Communications of the ACM*, 40(3):77–87, 1997.

[9] D. Logan. Understanding and using taxonomies. Technical report, Gartner, 2001. Resource ID: 330280.

[10] T. Molenaar. 'Intelligente documenten' vormen basis voor efficiëntie *('Intelligent documents' are the base for efficiency). Computable Business Review*, 4:30–33, 2004.

[11] H.-L. Ong, A.-H. Tan, J. Ng, H. Pan, and Q.-X. Li. Organizing and personalizing intelligence gathering from the web. *International Journal of Intelligent Systems in Accounting, Finance & Management*, 11:9–21, 2002.

[12] S. Shearin and H. Lieberman. Intelligent profiling by example. In *Proceedings of the International Conference on Intelligent User Interfaces (IUI)*, January 2001.

[13] S. Spek, K.-J. van Dorp, E. Mathijsen, Th. de Haas, and J. van den Herik. Advanced information search within a research-based multinational. In Tom Heskes, Peter Lucas, Louis Vuurpijl, and Wim Wiegerinck, editors, *Proceedings of the 15th Belgium-Netherlands Conference on Artificial Intelligence*, pages 283–290, 2003.

[14] K. van Rijsbergen. *Information Retrieval*. Butterworths, 1979.

[15] B. Vlugt. Ingenieurs alleen maken geen kennisland *(Engineers on themselves do not make a knowledge landscape). Computable Business Review*, 4:22–29, 2004. Interview with Ruud Smits.

[16] M. Weggeman. *Kennismanagement, inrichting en besturing van kennisintensieve organisaties*. Scriptum Management, Schiedam, The Netherlands, fourth (modified) print (2001) edition, 1997.

[17] S. Wright. Personalisation, how a computer can know you better than yourself. In *Proceedings of the Third Multimedia Systems Conference*, Southampton, January 2003.