On Current Technology for Information Filtering and User Profiling in Agent-Based Systems, Part I: A Perspective*

Sander M. Bohté William B. Langdon Han La Poutré
CWI, Centre for Mathematics and Computer Science
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands
S.M.Bohte@cwi.nl W.B.Langdon@cwi.nl Han.La.Poutre@cwi.nl

January 2000

Abstract

Several current techniques and methods in information filtering and profiling are surveyed, including state-of-the art technology, various techniques currently used by large businesses, and the academic state-of-the-art projects. Given the simplicity of the techniques currently applied in the field, the further development and application of technology currently available in AI and algorithmics will yield significant improvements in both filtering and profiling results.

1 Introduction

The terms "Information Filtering" and "profiling" are widely used. Here "Information Filtering" will refer to computer software systems which

- split (usually large) data streams into useful and not useful components and direct the useful to interested users. Of particular interest are systems which recognize their users are different and split the data stream into separate (possible overlapping) streams which are directed at distinct users or groups of users. Typically data in the streams is composed of atomic documents, which are treated individually. They are not split and are passed entirely to the relevant user(s) (if any). These systems are passive, data flows through them.
- In contrast intelligent information retrieval systems actively search for relevant information. Typically by searching the Internet. Typically

*This is a first, preliminary report in a sequence of short reports (Parts) on Current Technology for Information Filtering and User Profiling in Agent-Based Systems.

each search is in response to a request generated autonomously by a person.

The "Profiling" comes from marketing and credit risk assessment. It is a term for many techniques that group together people and assign a label to them. Grouping is often on the basis of basic social-economic data, e.g. gender, age, salary, zip code. Labels have consequences like "try and sell her a car", "don't give him credit". Here "intelligent profiling" refers to (upcoming) advanced computer techniques which perform classification automatically attempting to learn user preferences on-line and quickly. These are much more sophisticated and automatically and dynamically generate profiles specific to individual customers rather than simply assigning them to predefined, often binary ("yes", "no") categories. Of particular interest is the combination of profiling and information filtering, so only relevant information is presented to users and presented in a timely fashion. In the future, they may be used with software agents as well as people.

Typically in information filtering: the information need is (relatively) *stable* and *specific*, but the information sources are *dynamic* and *unstructured*. I.e. well established existing techniques which enable ad-hoc queries of structured data are excluded, e.g. relational databases, library catalogues. As such, information filtering is essentially a two way street: the information need can be established in terms of preferences (user profiling), and the information sources can be clustered and matched (information profiling).

Profiling can be subdivided into active and passive profiling. Most systems currently use passive profiling, where by passive we mean that a user is, given a set of attributes, classified as belonging to one or multiple predetermined clusters. Information is parsed in the same way, and thus a match can be established between the user's information need and the information sources. The pre-set clusters usually result from the application of data-mining techniques on the collected user-data and a subset of representative information sources and as such are relatively static. Examples of such applications are discussed in section 3.

In contrast, filtering and profiling methods based on agent technology aim to enable active profiling by establishing unique user-profiles and matching such profiles to (pre-processed) information sources. Some research is currently done to "learn" the user's profile in an unobtrusive and useful manner, and also adaptively learn "shifts" in a user's profile. The same active adaptation could also be applied to selecting information sources: agents could be assigned the task of adaptively finding information sources; an agent representing the user preferences could then determine which "filtering" of what sources is most relevant to the user.

In section 2 we present some core-techniques used in current information filtering systems, section 3 states several techniques currently in commercial use. Section 4 lists some of the important (academic) projects currently

underway and the means by which they aim to achieve more refined information filtering and user profiling. Although in a sense agent-mediated electronic commerce can also be considered a form of information filtering (filtering only potentially interesting products), a survey of this field is beyond the scope of this report, and we refer to the excellent survey written by Guttman, Moukas and Maes from MIT [Guttman *et al.*, 1998].

The purpose of this report is to give a first introductory and exploratory tour through the technology field; it is certainly not meant to give an exhaustive or complete overview. In later reports (Parts), we will discuss specific subjects and applications in more depth and completeness.

2 Profiling and Filtering Technologies

Most information filtering systems are based on a number of key-techniques used to describe information, create a user profile and create the interaction and filtering needed for a useful system. It is these technologies that are implemented with ever increasing refinement and more complex variations in the ongoing research we discuss in section 4. Here we sketch the most used technologies.

2.1 Key filtering technologies

2.1.1 Keyword vectors

Keywords are the most popular way of representing documents and are also used to represent user-profiles. Most representations are based on a standard information retrieval technique called weighted vector representation ([Salton and Buckley, 1987]). Of a (pre-processed) document, the text is decomposed into its keywords, which are weighted relative to importance (as could be established by a frequency count), and thus compose the keyword-vector. Document similarity and document distance to a preferred profile-vector can be easily obtained by comparing the respective vectors with for instance k-Nearest-Neighbor algorithms [Shepard, 1968]. The same technique can also be used to obtain user-profiles.

The weighting formula often used, as suggested in [Salton and Buckley, 1987] is called the "term frequency times inverse document frequency " ("tfidf") and calculates the weight for each keyword as:

$$W_k = H_c \cdot T_f \cdot idf_k \tag{1}$$

where T_f is the frequency of the keyword in the current document (*term frequency*), H_c is a constant related so information like the title (a potential pre-classification biasing), and idf_k is formally defined as:

$$idf_k = log(\frac{N}{df_k}) \tag{2}$$

N is the number of documents that have already been retrieved by the system and df_k is the frequency of the keyword in the entire collection of documents (*document frequency*). User profiles can be obtained by determining (clusters of) document vectors that are indicative for the type of information of interest to the user.

2.1.2 N-grams

A recent tool for representing documents and/or user-profiles are n-gram distributions. An n-gram is a sequence of n letters. Typically n is at least three. For each n and size of alphabet there are a finite number of letter sequence of length n and thus a fixed number of n-grams. A text can be converted to an n-gram distribution by counting the number of times each possible n-gram appears within the text. Note n-grams overlap. Usually text is converted to uppercase and all non-alphabetic characters are replaced by spaces and multiple spaces are reduced to a single one. The main benefit of N-grams lies in the fact that they are less sensitive to spelling-errors and that the (large!) n-gram vector also incorporates more of the document structure as compared to keywords. Allegedly the US National Security Agency (NSA) uses these techniques to automatically classify documents obtained via their Echelon network: they have a patent on a special n-gram weighting algorithm [Damashek, 1995]. More accessible, an adaptive information filtering system based on weighted n-grams has been developed by Daniel Tauritz [Tauritz and Sprinkhuizen-Kuyper, 1999].

2.1.3 Hyperlink structures

Specifically for documents with linked structures, such as web-pages, graph-like representations can be extracted, mapping out the relationships between documents, and between words near links to other documents. Such structure can be exploited to filter web-pages into different categories. Work of this kind is actively being pursued in the group of Tom Mitchell of Carnegie Mellon [Craven *et al.*, 1999], `http://www.cs.cmu.edu/~webkb/`, and at for instance the Austrian Research Institute for AI [Fuernkranz, 1999].

2.1.4 Collaborative and economic-based filtering

Collaborative (or social) filtering utilizes feedback and ratings from different users to filter out irrelevant information. The information interesting for a user is gathered "on the fly" by using the opinions of other users with similar interest. Economic-based filtering augments this idea with a cost-benefit analysis on behalf of the user. It takes into consideration parameters like the price of a document and its cost of transmission when making filtering decisions. Examples of such systems include PHOAKS [Terveen *et al.*, 1997], GroupLens [Resnik *et al.*, 1994] and Ringo [Shardanand and Maes, 1995].

2.1.5 Data-mining techniques

Data-mining techniques can be employed to find similarities between data-entries, and thus inferring that the profile of a given user might be very close to the profile of some other users. Thus correlating the current customer to previous users (people-to-people correlation, e.g. you are like this type of customer who typically likes...) or to these previous users' interests (item-to-item correlation, e.g. this item you are considering is very much like these items...) allows companies to present a customer with information that he or she is likely to be interested in. This is a simple but very popular application of information filtering (see section 3).

2.2 Key user-modeling techniques

User modeling can be defined as the effort to create a profile of the user's interests and habits. User Modeling systems differ in the way they acquire, use and represent a profile. Profiles could be acquired or generated in a variety of ways:

1. By explicit modeling by humans:
    - By direct user interviews and questionaires
    - By "knowledge engineers" using user stereotypes.
    - Rule-based profiles, where the users specify their own rules in the profile, rules that control the behavior of the model.
2. By automated software techniques:
    - Machine learning techniques like inference, induction and classification, where the modeler tries to identify certain patterns in the user's behavior.
    - Profile building by example, where the user provides examples of his/her behavior and the modeling software records them.

At the moment, (1) is much further developed and significantly more applied than (2), which is in its development phase. For some of the currently-in-use profiling systems we will discuss what profiling systems they employ in section 3.

2.3 Key agent technologies

Agent technologies are based on methods to create systems of (small) software-programs that through interaction with each other and with the outside world are able to achieve a computational goal in an essentially distributed computing paradigm. For profiling and information purposes, these agent models are characterized by the following notions and potentials:

- a user's preferences are incorporated in one or multiple agents acting on behalf of the user as the actual information filter.
- information can be acquired from a multitude of sources, either by direct inquiry on search engines, or by delegating a request for information to agents specialized in information retrieval (e.g.: Amalthea [Moukas, 1997]).
- in more extended systems, ecologies of agents may be created: set of user agents interacting with sets of information retrieval agents. In e.g. [Moukas, 1997], such systems can subsequently enhance their effectiveness by using economic models for assigning credit to (un)successful operations thus "learning" what interaction realize sufficient pay-off.
- learning in such systems may be incorporated by machine learning algorithms, such as reinforcement learning (e.g. [Sycara, 1999, Sycara, 1998]), or by techniques such as evolutionary algorithms (e.g. [Moukas, 1997], where unsuccessful agents are replaced by newly created agents).

These notions and potentials form important ingredients of current and future agent technologies for information filtering and profiling applications.

3 Information Profiling Instances

In this section, we consider current information profiling applications in businesses. An extensive compilation of currently used techniques in several large (web) businesses can be found in [Schafer *et al.*, 1999]. In this section, we highlight some of the important business instances and aspects, which are not mentioned above.

A number of companies sell specialized software for information-mining. Information on the actual methods they employ is not readily available, but keyword approaches or explicit modeling usually play an important role. We limit ourselves here to mentioning a few of such companies: Muscat (`http://www.muscat.com`), Netperceptions, a spin-off of the Univerity of Minnesota information filtering effort (`http://www.netperceptions.com/home/`), and Frictionless, spin-off of the MIT Agents Technology lab (`http://www.frictionless.com/`).

Some companies on internet offer filtering of (product) information to their users. Examples are Amazon.com (`http://www.amazon.com`), CDNOW (`http://www.cdnow.com`), eBay (`http://www.ebay.com`), and Moviefinder.com (`http://www.moviefinder.com`). The techniques they employ are rather simple. In general, two main techniques are most often used: on item-to-item correlation and people-to-people correlations. For people-to-people correlations, the correlation is calculated for items bought by the same person: if book A and book B are often bought by the same

customer, a customer considering book A will be recommended to also look at book B. For item-to-item correlation, an attempt is made to classify items based on their content or type of product, and recommend similar items to a customer.

These recommendations are thus made based on information obtained through data-mining results, which is then matched to the current customer's profile (e.g. current and previous product interests). To aide the customer in deciding on what item to purchase from many similar items, a rating is often collected through the feedback responses of previous buyers, thus employing simple collaborative rating and reviewing.

For a further description of currently used techniques in several large (web) businesses, we refer to [Schafer *et al.*, 1999].

4 Information Filtering and Profiling Projects

Most current projects that aim at information filtering are centered around agent-based solutions, mostly in order to provide a more personalized profile thus hopefully providing the user with a greater overall sense of accurate filtering. In this section we give a summary of several current and past projects on information filtering. In section 4.1, we state some approaches which are more or less completed, or which have already yielded significant results. In section 4.2, we summarize several major current efforts under way, aimed at either extending current technologies or creating new ones.

4.1 Mature projects

Various academic projects are more or less mature in the sense that prototype research software has been created and current work mostly deals with optimizations.

The group of Kohonen, at the Helsinki University of Technology (HUT) has developed a mature application based on self-organizing neural networks [Kaski *et al.*, 1998]. This application organizes documents into large multi-dimensional maps based on keyword vectors. The resultant maps allow one to easily find documents of similar content. Based on this work, alternative adaptive self-organizing schemes have been explored [Schuemie and van den Berg, 1999]. The application could be extended to include actual information filtering by including a user-profile to automatically select information, but in the current state, a user has to actively engage the system with queries to find related documents.

Started in 1992, the group of John Riedl at the University of Minnesota has been working on the GroupLens project, aimed at exploring automated collaborative filtering [Resnik *et al.*, 1994]. The results obtained on netnews filtering have led to the formation of a company to commercially exploit the technology, whereas the current research is more concerned with automated

collaborative filtering and classification of non-textual information such as movies and music.

An MIT effort which can be classified as "mature" pertains to the Amalthea project [Moukas, 1997]. Based on an ecology of agents, it has been shown to enable reasonably good dynamic information filtering based on adaptive user modeling through GA techniques operating on information filtering agents and effective information retrieval through GA based information discovery agents. The work has resulted in a software prototype, but appears to be currently on hold due to diverting interests.

MIT's Letizia project, started in 1995, attempts to suggest to a user some links to follow next when surfing the World Wide Web [Lieberman, 1995]. From the user's current position, the system conducts a concurrent breadth first search, results. The system maintains a frequency list of keywords previously encountered, and compares possibly recommendable documents against this list, thus attempting to match links to the user's interest. An extension of this system is being developed in the MIT project "Let's Browse", which is an experiment aimed at building agents capable of assisting a group of people in browsing, by suggesting new material likely to be of common interest (provided by the Letizia framework).

The Stanford Digital Library project FAB has resulted in a prototype system that aims to learn to browse the Internet on behalf of a user based on the observation day-to-day use. FAB recommendations are based on a combination of collaborative filtering and content-based filtering [Balabanovic and Shoham, 1997]. The content-based filtering works by recommending texts that fit the user-profile via a weighted keyword comparison. The collaborative filtering system recommends texts that other similar users have liked. To compute the similarity between users, a nearest-neighbor search is performed on users with similar previous ratings of texts. The system has been operational in several versions since 1994.

4.2 Current ongoing projects

In this section, we briefly survey current academic efforts related to improving information filtering. The most significant projects aimed at information filtering are currently in progress at the Software Agents Lab of MIT and the Agent Lab at Carnegie Mellon University., concerning workflow and scheduling simulations. At Carnegie Mellon University (CMU), the RETSINA project aims to develop a framework for distributed collections of intelligent software agents that perform goal-directed information retrieval and information integration in support of a variety of decision making tasks [Paolucci *et al.*,]. Being a framework, the system is under continual development, but actual applications have been created based on the system, specifically for use in workflow situations where the systems retrieves relevant information for a user based on the task at hand [Shehory *et al.*, 1999].

The Pleiades project at CMU concerns scheduling problems and is aimed at improving agent learning techniques relating to learning user profiles, as well as negotiation techniques to enable multiple agents belonging to the same user to find optimal multi-issue solutions (e.g.: finding calendar time for meetings and other requirements) [Mitchel *et al.*, 1994].

A fairly specific application of information filtering is pursued in MIT's Butterfly project [Van Dyke *et al.*, 1999]. Started in 1998, this project attempts to monitor Internet Relay Channels (or chat-boxes) for discussions which might interest the user. The current version is based on frequency-weighted keyword vectors to represent the user's interests, and a semi-interactive program to create frequency-weighted keyword vectors from a limited number of IRC channels. Current work is focussed on extending the system to a true agent-based one, where multiple agents can monitor different sets of IRC channels and interact with user-agents to determine whether or not an IRC channel is of interest.

A system similar to MIT's Amalthea is being devised at IBM: Grand Central Station (GCS). This system also consists of both an information filtering part, that matches information against the profiles of individual users and an information discovery part that constantly gathers and summarizes new information. In the Netherlands, the Information Retrieval and Information Systems research line in the Computing Science Institute at Nijmegen has worked for a few years on Internet based information retrieval systems. While initially active in agent based approaches [Wondergem *et al.*, 1997], recent interest has tended to focus on improved syntax of *index expressions* (keyword, concept names, or denominations of attribute values *plus* the relationships between them) for user queries of the WWW [Wondergem *et al.*,]. Other projects that have recently been started and aim to develop agent-based information filtering tools are the collaborative EU project *Select*, and the more application driven, current development of the agent infrastructure at Tryllian, Gossip.

5 Summary

Sections 2, 3 and 4 have described several key text filtering technologies, major existing Internet based commercial users of profiling techniques and some key research projects on information filtering and profiling. Individualised profiling is only in a few applications done other than by broad categorization based on categories developed off-line. Agent systems appear to offer a possible way in which World Wide Web based business could reap the potential rewards of interactively responding to individual customers in real time.

References

[Balabanovic and Shoham, 1997] M. Balabanovic and Y. Shoham. Fab: content-based, collaborative recommandation. *Communication ACM*, 40:66–72, 1997.

[Craven *et al.*, 1999] M. Craven, D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam, and S. Slattery. Learning to construct knowledge bases from the world wide web. *Artificial Intelligence*, pages 1–58, 1999.

[Damashek, 1995] Marc Damashek. Method of retrieving documents that concern the same topic. *United States Patent*, pages 1–19, 1995.

[Fuernkranz, 1999] J. Fuernkranz. Exploiting structural information for text classification on the www. In David J. Hand, Joost N. Kok, and Michael R. Berthold, editors, *Advances in Intelligent Data Analysis, Third International Symposium, IDA-99*, volume 1642 of *LNCS*, pages 487–497, Amsterdam, The Netherlands, 9–11 1999. Springer-Verlag.

[Guttman *et al.*, 1998] R. Guttman, A. Moukas, and P. Maes. Agent-mediated electronic commerce: A survey. *submitted*, pages 1–8, 1998.

[Kaski *et al.*, 1998] S. Kaski, T. Honkela, K. Lagus, and T. Kohonen. Websom–self-organizing maps of document collections. *Neurocomputing*, 21, 1998.

[Lieberman, 1995] Henry Lieberman. Letizia: An agent that assists web browsing. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 1995.

[Mitchel *et al.*, 1994] T.M. Mitchel, R. Caruana, D. Freitag, J McDermott, and D. Zabowski. Experience with a learning personal agent. *Communication ACM*, 37:91, 1994.

[Moukas, 1997] A. Moukas. Amalthaea: Information filtering and discovery using a multi-agent evolving system, 1997.

[Paolucci *et al.*,] M. Paolucci, D. Kalp, A. Pannu, O. Shehory, and K. A Sycara. Planning component for retsina agents. In M. Wooldridge and Y. Lesperance, editors, *Lecture Notes in Artificial Intelligence, Intelligent Agents VI.* Springer Verlag.

[Resnik *et al.*, 1994] P. Resnik, N. Iacovou, M. Sushak, P. Bergstrom, and J. Riedl. Grouplens: An open architecture for collaborative filtering of netnews. In *Proc of CSCW'94*, 1994.

[Salton and Buckley, 1987] G. Salton and C. Buckley. Text weighting approaches in automatic text retrieval. *Cornell University Technical Report 87-881*, 1987.

[Schafer *et al.*, 1999] J.B. Schafer, J. Konstan, and J. Riedl. Recommender systems in e-commerce. In *Proc. ACM Conf on E-Commerce 1999*, 1999.

[Schuemie and van den Berg, 1999] M. Schuemie and J. van den Berg. Information retrieval systems using an assosiative conceptual space and self-organizing maps. In E. Postma and M. Gyssens, editors, *BNAIC 1999*, pages 91–97, 1999.

[Shardanand and Maes, 1995] U. Shardanand and P. Maes. Social information filtering: Algorithms for automating 'word of mouth'. In *Proc. CHI-95 Conf.*, 1995.

[Shehory *et al.*, 1999] O. Shehory, K. Sycara, G. Sukthankar, and V. Mukherjee. Agent aided aircraft maintenance. In *Proceedings of Third International Conference on Autonomous Agents.*, 1999.

[Shepard, 1968] D. Shepard. A two-dimensional interpolations function for irregularly spaced data. In *Proceedings of the 23rd Nactional Conference of the ACM.*, pages 517–523, 1968.

[Sycara, 1998] K. Sycara. Levels of adaptivity in systems of coordinating information agents. In M. Klusch and G. Weiss, editors, *Cooperative Information Agents II.* Springer Verlag, 1998.

[Sycara, 1999] K. Sycara. In-context information management through adaptive collaboration of intelligent agents. In Matthias Klusch, editor, *Intelligent Information Agents: Cooperative, Rational and Adaptive Information Gathering on the Internet.* Springer Verlag, 1999.

[Tauritz and Sprinkhuizen-Kuyper, 1999] Daniel R. Tauritz and Ida G. Sprinkhuizen-Kuyper. Adaptive information filtering algorithms. In David J. Hand, Joost N. Kok, and Michael R. Berthold, editors, *Advances in Intelligent Data Analysis, Third International Symposium, IDA-99*, volume 1642 of *LNCS*, pages 513–524, Amsterdam, The Netherlands, 9–11 1999. Springer-Verlag.

[Terveen *et al.*, 1997] L. Terveen, W. Hill, B. Amento, D. McDonald, and J. Creter. Phoaks a system for sharing recommendations. *Communication ACM*, 40:59–62, 1997.

[Van Dyke *et al.*, 1999] Neil W. Van Dyke, Henry Lieberman, and Pattie Maes. Butterfly: A conversation-finding agent for internet relay chat. In *Proceedings of the 1999 International Conference on Intelligent User Interfaces*, pages 91–97, 1999.

[Wondergem *et al.*,] B.C.M. Wondergem, P. van Bommel, and T. van der Weide. Nesting and Defoliation of Index Expressions for Information Retrieval. *Knowledge and Information Systems.* To Appear.

[Wondergem *et al.*, 1997] B.C.M. Wondergem, P. van Bommel, T.W.C. Huibers, and Th. van der Weide. Towards an Agent-Based Retrieval Engine. In J. Furner and D.J. Harper, editors, *Proceedings of the 19th BCS-IRSG Colloquium*, pages 126–144, Aberdeen, Scotland, April 1997. Robert Gordon University.