Opportunities for Electronic Commerce in Agent-Based Information Discovery

B.C.M. Wondergem, P. van Bommel,
T.W.C. Huibers, and Th.P. van der Weide
Computing Science Institute, University of Nijmegen
Toernooiveld 1, NL-6525 ED, Nijmegen, The Netherlands
tel: +31 24 3653147, fax: +31 24 3553450
E-mail: bernd@cs.kun.nl

Published as: B.C.M. Wondergem, P. van Bommel, T.W.C. Huibers, and Th.P. van der Weide. Opportunities for Electronic Commerce in Agent-Based Information Discovery. In, F. Griffel, T. Tu, and W. Lamersdorf, editors, *Proceedings of the International IFIP / GI Working Conference on Trends in Distributed Systems for Electronic Commerce (TrEC98)*, pages 126 – 136, Hamburg, Germany, June 1998.

Keywords: Information Discovery, Electronic Commerce, Intelligent Agents, Middle-Agents, Information Brokerage, and Multi-Agent Systems.

Abstract

This article investigates the connection between Electronic Commerce (EC) and Information Discovery (ID). ID is the synthesis of distributed Information Retrieval and Information Filtering, filled in with intelligent agents and information brokers. Currently, no link exists between EC and ID. We argue that this link consists of a cost model for ID. We therefore propose several (types of) cost models, which enable application of EC to the whole of ID. This is illustrated with examples.

1 Agent-Based Information Discovery

The quest for relevant information has given rise to two major plans of attack: Information Retrieval (IR) (see [Rij75]) and Information Filtering (IF). IR and IF, which are elaborately compared in [BC92], have received great attention over the past decades. The enormous expansion of the Internet guarantees a non-declining interest in the near future.

Information Discovery (ID) (see [WBHW97]) is the synthesis of IR and IF. In ID, a networked population of users having dynamic information needs is considered. The information needs are to be satisfied with documents out of dynamic information sources. To facilitate this process, information brokers act as intermediaries between users and sources. Brokers compare the wishes of users with the available information in the sources. Figure 1(a) sketches the ID paradigm at a conceptual level:

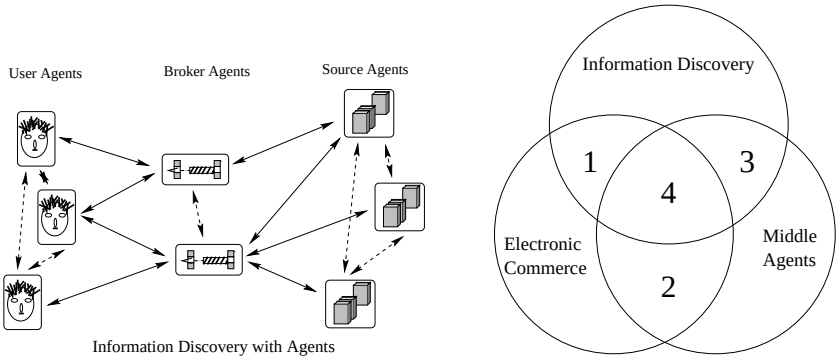

Figure 1: (a) Conceptual ID Paradigm (b) Three Research Areas Combined

Since agent technology (see e.g. [WJ95]) suits the ID paradigm perfectly (especially communication, intelligence, and proactiveness - see [WBHW98]), the entities of the ID paradigm in figure 1(a) are seen as agents: user agents, broker agents, and source agents, respectively. User agents within ID are focussed on in [Sim97]. The community of agents in ID forms a multi-agent system. The arrows in figure 1(a) represent communication. Note that information brokers are true *middle-agents* since it is not possible for users to communicate directly with information sources.

The multi-agent system forms a networked 'market' of ID agents which offer (to sell) documents and pose requests for (buying) documents. This does not, however, allow *Electronic Commerce* (EC) (see e.g. [Zwa96]) to be directly applicable to ID. The main obstacle is that ID is traditionally not focussed on the principle of buying and selling, but on *matching* user queries with available documents in order to deliver sets of *relevant* documents to users. In [MA95] suggestions are made towards better use of EC on the Internet.

The missing link between EC and ID appears to be a *cost model for ID*. If this cost model for ID is directly based on money, other ID features can

no longer be dealt with properly. This is because these ID features are about key concepts of ID, e.g. documents and user queries, and not about money. Therefore, it is necessary to define cost models for ID in terms of concepts from within ID. This ensures applicability of EC to the whole of ID. The goal of this article is to analyse the possibilities for such cost models.

We propose several cost models that not only enable matching to be treated from an EC perspective, but also more advanced ID features, such as query expansion, profile adaptation, and relevance feedback. This implies that ID is made accessible for EC, resulting in a large and tentative new area of EC-based research.

Figure 1(b) provides a Venn-diagram for the three research areas combined in agent-based ID: ID, Middle-Agents, and EC. Region 1 represents a client-server information market, region 2 corresponds to mediated markets, region 3 stands for mediated ID, and region 4 expresses the main focus of this paper: a *mediated information market*.

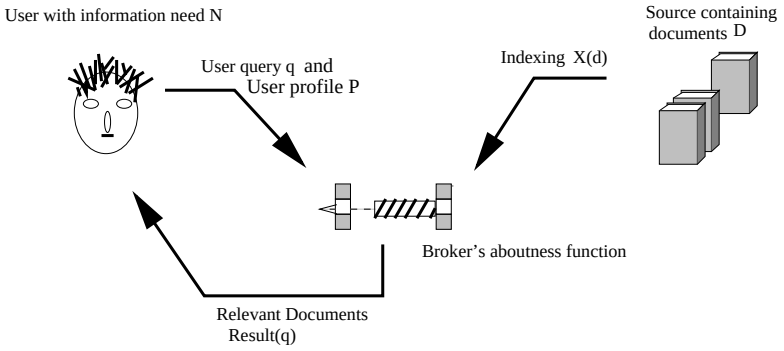

Figure 2: Key Concepts of ID

The basic preliminaries of ID are presented in figure 2). On the left hand side, we find users having information needs, denoted by N. Information needs are divided into a short term need, formulated in query $q \in L$, and long term needs, formulated in a user profile $P \subseteq L$. Here, L denotes the descriptor language, e.g., keywords or index-expressions.

On the right hand side, sources are shown that each support a set of documents D, and an indexing or characterisation function $\chi : D \mapsto \wp(L)$, delivering a set of descriptors as characterisation of a document.

The central part of figure 2 depicts information brokers which implement a matching function, also aboutness function, $\texttt{about} : D \times L \mapsto [0..1]$, that establishes the match (degree of aboutness) between document d and request q. For reasons of simplicity, we assume that the elements of the user profile P are also denoted by q. Aboutness approximates the extent to which document d is relevant to the information need N. Documents which have a high degree of aboutness constitute the result set of a query and are rendered: $\texttt{result}(D, q) = \{d \in D | \texttt{about}(d, q) > \rho\}$, for some threshold $0 \leq \rho < 1$.

The rest of this paper is organised as follows. The next section is the main focus of this paper and describes a number of cost models for ID. Section 3 touches upon the applicability of EC to some additional ID features. Section 4 provides concluding remarks and suggestions for further research.

2 Cost Models for ID

As stated in the introduction, the missing link between EC and ID is a *cost model for ID*. The cost model should define the cost of the items and services offered. Cost models for ID can be defined in a variety of ways, several of which are examined below.

Within ID, two different categories of cost models are possible: *intrinsic* and *user relevance* cost models. In intrinsic models, cost is a direct feature of the item or service sold. In user relevance models, cost is defined in terms of the mismatch between user preferences and offered characteristics. Here, offered features are source documents (see section 2.2.1) and broker matching services (see section 2.2.2).

We provide *type definitions* of cost functions which can be instantiated in numerous ways. The result type of cost functions is C, which stands for cost. Instantiations of type C are transaction cost, processing cost, reading effort, etc. The input types depend on the point of view with which the cost function is defined. Accordingly, the different possible points of view of ID lead to different cost models. It is, of course, possible to combine several cost functions. For instance, users can be charged for both broker matching services and source documents.

2.1 Intrinsic Cost Models

Intrinsic cost models for ID can be defined from three different points of view: sources, brokers, and users, since each type of agent can charge for its service or items offered. Intrinsic cost models do not take user preferences into account. Therefore, in order to ensure proper ID functionality, they should be used as a layer on top of ID-aboutness. This means that EC is applied to the pre-computed set of relevant documents: $\texttt{result}(D, q)$.

1. **Source: Documents and Characterisation**

   Sources offer documents $d \in D$ and their characterisations $\chi(d) \in \wp(L)$. Users can be charged for documents since they are in need of information. In addition, brokers can be charged for documents needed as, e.g., test collection to tune the matching function.

   First of all, cost models for sources can be defined in terms of intrinsic features of documents. In ID, these features are called *Situational Factors* (SF). SF's do not describe the informational content of documents, which is formulated in $\chi(d)$, but additional characteristics such as size, date of creation, number of pictures, etc. The input type of a source cost model can thus be (the SF's of) D:

$$\texttt{cost} : D \mapsto C$$

Example instantiations are:

$$\texttt{cost}(d) =^{\texttt{def}} \texttt{size}(d), \qquad \texttt{cost}(d) =^{\texttt{def}} \texttt{age}(d)$$

In addition, the cost for documents could, of course, be a constant. This constant could, for instance, stand for administration cost.

Second, the characterisation function χ also forms a basis for a cost model for sources:

$$\texttt{cost} : \wp(L) \mapsto C$$

The idea is that better document characterisations lead to better retrieval performance since matching is more detailed. More elaborate characterisations, however, are more costly to compute. This is reflected in the following examples:

$$\texttt{cost}(d) =^{\texttt{def}} |\chi(d)|, \qquad \texttt{cost}(d) =^{\texttt{def}} |L|$$

If communication among sources is considered, each source can charge a cost that reflects the relative quality of its characterisation function with respect to the quality of the others'. Consider a set of sources S. Then for source $s \in S$:

$$\texttt{cost}_s(d) =^{\texttt{def}} \frac{|\chi_s|}{|\bigcup_{s' \in S} \chi'_s|}$$

Examples of sources charging cost are digital libraries and digital publishing. Combinations of these cost models are possible. For instance, cost could be charged that depends on both the size of the document and the quality of its characterisation.

2. **Broker: Matching Function**

   Brokers offer aboutness functions with which they compute the set of relevant documents $\texttt{result}(D, q)$. Brokers could charge for the computation of this set. In our opinion, it will not take long before search engines (brokers) on the Internet such as AltaVista[1] and Infoseek[2] will start charging for their services.

   ID brokers perform 2-way matching: both users and sources can ask for broker services. In general, the agent requesting a service also pays the costs. Thus, both users and sources can be charged by brokers. The costs a broker charges are based on the broker's input and features of its matching function:

[1] Available from: http://www.altavista.digital.com/
[2] Available from: http://guide.infoseek.com/

$$\texttt{cost} : D \times L \mapsto C$$

The cost a broker charges can be relative to the number of resulting relevant documents, the size of the query, or the amount of documents to be processed:

$$\texttt{cost}(D,q) =^{\texttt{def}} |\texttt{result}(D,q)|, \qquad \texttt{cost}(D,q) =^{\texttt{def}} |q|,$$
$$\texttt{cost}(D,q) =^{\texttt{def}} |D|$$

In practice, brokers may charge more during peak hours. In addition, communication between brokers offers them the possibility to adjust their cost based on the costs of other brokers. Criteria could be, for instance, efficiency, quality, or the amount of domain knowledge used in the matching. In [Hui96], broker matching functions are described on an axiomatic level. This allows an embedment relation on matching functions to be defined. This relation could serve as a measure of quality.

3. **User: Query and Profile**

   Although users mostly are at the paying end, they can provide their query $q \in L$ and profile $P \subseteq L$ for use to brokers and sources and charge accordingly. Brokers building an overview of user interests and sources desiring to perform IF are interested in this information. The input type therefore is L, the descriptor language.

   The result type of the user cost model could be documents, viz. free information for the user, or the endurance that broker services will stay available to the user.

$$\texttt{cost} : L \mapsto C$$

   Examples illustrating the idea that more information costs more:

$$\texttt{cost}(q) =^{\texttt{def}} |q|, \qquad \texttt{cost}(P) =^{\texttt{def}} \textstyle\sum_{q \in P} \texttt{cost}(q)$$

   We do not advocate that the situation where users charge brokers and sources is the most common. We merely show that this situation can be described in terms of user cost models.

2.2 User Relevance Cost Models

Opposed to intrinsic cost models, user relevance cost models take user preferences into account. These preferences can concern documents and broker matching functions. As shown in section 3, they can also serve as basis of additional ID features.

2.2.1 Document Relevance

ID is concerned with the deliverance of relevant documents to users. Relevance of documents is always relative to user interests. Since relevance is a user-oriented criterion, user relevance cost models will also be user-centered. The basic ID mechanism to approximate relevance is aboutness, which is the match between query and document characterisation. We show how user relevance cost models can be defined in terms of aboutness.

Aboutness only concerns the informational content of documents. A more elaborate approximation of relevance also takes user preferences on additional document characteristics (as formulated in situational factors) into account.

1. **Aboutness**

   The cost for a user to process a document is inversely related to the document's relevance to the user query. The more relevant the document, the lower the processing costs. A less relevant document costs the user more effort to process, since it is harder to link with the information need.

   This leads to user determined costs for documents.

   $$\texttt{cost} : D \times L \mapsto C$$

   An example instantiation:

   $$\texttt{cost}(d, q) =^{\texttt{def}} (1 - \texttt{about}(d, q))$$

   Another approach would be to charge users more for documents that are more relevant. The difference with the abovementioned approach is that, then, the cost of a document would exist next to its relevance, and not, as in our case, be defined in terms of its relevance.

2. **Situational Factors**

   A more direct notion of cost for documents can be defined in terms of the SF's of the document. With each document a set of SF's is associated. In addition, user preferences regarding SF's are stated in the query. As with aboutness, the mismatch between the two forms a cost model:

   $$\texttt{mismatch} : D \times L \mapsto C$$

   The total mismatch is the sum of mismatches of individual SF's. Assuming a `mismatch'` function for individual SF's $f \in SF$, the overall mismatch can be defined as:

   $$\texttt{mismatch}(d, q) =^{\texttt{def}} \frac{\sum_{f \in SF} \texttt{mismatch'}(f_d, f_q)}{|SF|}$$

   Some SF's may be not or partially specified. Mismatches should then be less severe, expressing the fact that the particular SF is not of great importance to the user.

3. **Aboutness & Situational Factors**

   The abovementioned approaches can be combined to obtain a complete user relevance cost model. Assume that the cost based on aboutness is formulated in `Topic` and that `SitFac` expresses the cost related to SF's. Then, the most general way to combine them is (with $\alpha \in [0,1]$):

$$\texttt{cost}(d,q) =^{\texttt{def}} \alpha \times \texttt{Topic}(d,q) + (1-\alpha) \times \texttt{SitFac}(d,q)$$

   In this way, all features of documents with respect to user preferences are taken into account.

2.2.2 Broker Matching Services

A broker's matching function can also be subject to user preferences. This idea evidently connects to work of Huibers and Denos, who express aboutness at the level of postulates in [HD95]. User preferences on these postulates are exploited to obtain qualitative document rankings. The underlying idea of ranking with respect to user preferences on aboutness postulates can be used to define cost models for broker services.

In order to accomplish this, the aboutness functions are described in terms of postulates. User preferences on individual postulates are then lifted to complete aboutness functions. In this way, a user defined preference relation $\prec$ is obtained, such that $b \prec b'$ means that the user prefers the aboutness function of broker b above the one of b'. The according cost model should satisfy

$$b \prec b' \quad \Leftrightarrow \quad \texttt{cost}_b < \texttt{cost}_{b'}$$

This ensures that more preferred brokers are cheaper and thus more easily contracted by the user.

3 ID Features from EC Perspective

In this section, we describe a number of ID features from the point of view of EC. That is, we elaborate on additional ID functionality in terms of concepts of EC, i.e. cost models. Since user relevance cost models take user preferences into account, they are also suitable for other ID tasks involving user preferences.

Some of the aspects hinge on cooperation among agents. In terms of EC, important aspects concerning cooperation are team formation and negotiation (see e.g. [SL95]).

3.1 User Agents

Since users often find it hard to formulate their information need properly in a query, the user query frequently consists only of a small number of descriptors. This leads to large result sets containing many irrelevant documents. Adding suitable terms to the query before firing it, *query expansion*, reduces this deficit,

since the query then forms a more correct and complete description of the information need. Suitable terms can be taken from the user's profile or from lexical-semantic relations (see e.g. [Voo94]).

In terms of EC, query expansion comes down to minimizing the costs users have to pay both brokers, using a cost model based on the size of the result set, and sources, using a cost model based on the documents rendered.

Furthermore, the distributed nature of ID makes it possible for users to expand their queries also with respect to profiles of other related users. In this way, users can cooperate, setting the context for team formation and negotiation.

In an analogous way, user profiles should formulate the long term interests as good as possible. An advantage of the synthesis of IR and IF in ID is that user queries can be used to adjust the profile. *Profile adaptation* can take place (1) at the time of the formulation of the user query, (2) at the time of obtaining the result set, and (3) at the moment of relevance feedback (see also next subsection).

3.2 Broker Agents

Relevance feedback is an instrument with which the user can explicitly remark on the relevance of the documents in the result set. The user can specify whether or not the rendered documents correspond to the information need. The broker can use this information to adjust its matching function. In this way, further results will better fit the user's information need. In terms of EC, this means that relevance feedback is an instrument to minimize user costs.

Brokers can perform *cooperative matching* to obtain better quality. Examples of this idea are MetaSearch[3] and MetaCrawler[4], using a hierarchy of broker agents and merging results. In terms of EC, this means the combined matching function is more competitive than the individual ones.

3.3 Source Agents

Source agents can send out samples of documents to create user interest. This user interest concerns both the topic of the sample documents and the source itself as a provider of high quality documents. In addition, sources can offer gifts for using the source's services. Also, sources can place tentative offers for users and brokers.

In terms of EC, this relates to *advertising*. It is used in ID to create user preferences for particular sources. These preferences are later used in cost models based on situational factors.

[3] Available from: http://www.metasearch.com
[4] Available from: http://www.metacrawler.com

4 Conclusions

We showed how to make Information Discovery accessible for concepts and techniques of Electronic Commerce, by providing the missing link between the two fields: cost models for ID. Although the proposed cost models are rather basic, they show how EC can be applied to ID. Further research is needed into more complex cost models. Those can, for instance, be based on other preferences, such as browsing preferences and direct preferences for sources. In addition, research should be conducted into applying additional concepts of EC to ID, in order to enlarge the cross fertilisation.

Our ideas are currently being implemented in the context of the PROFILE information filtering project (see [WSA+96] and [WBHW97]) of the University of Nijmegen, The Netherlands. To implement the ID paradigm with *real* agents, the JATLite agent testbed is used. Here, agents communicate using KQML ([FFMM94]) and are implemented in JAVA.

References

[BC92] N.J. Belkin and W.B. Croft. Information filtering and information retrieval: Two sides of the same coin? *Communications of the ACM*, 35(12):29–38, December 1992.

[FFMM94] T. Finin, R. Fritzon, D. McKay, and R McEntire. KQML - a language and protocol for knowledge and information exhange. In M. Klein, editor, *Proceedings of the 13th International Distributed Artificial Intelligence Workshop*, 1994.

[HD95] T.W.C. Huibers and N. Denos. A qualitative ranking method for logical information retrieval models. Technical Report RAP95-005, Groupe MRIM of the Laboratoire de Génie Informatique, Grenoble, France, August 1995.

[Hui96] T.W.C. Huibers. *An Axiomatic Theory of Information Retrieval.* PhD thesis, Department of Computer Science, Utrecht University, November 1996.

[MA95] Z. Milosevic and Bond. A. Electronic Commerce on the Internet: What is still missing? In *Proceedings of the 5th Annual Conference of the Internet Society, INET'95*, Honolulu, Hawaii, 1995.

[Rij75] C.J. van Rijsbergen. *Information Retrieval.* Butterworths, London, United Kingdom, 1975.

[Sim97] J. Simons. Using a semantic user model to filter the world wide web proactively. In Anthony Jameson, Cécile Paris, and Carlo Tasso, editors, *Proceedings of the sixth international Conference UM97.* SpringerWienNewYork, 1997.

[SL95] T.W. Sandholm and V.R. Lesser. Automated contracting among self-interested bounded rational agents. Technical report, Computing Science Department, University of Massachusetts, Amherst, 1995.

[Voo94] E.M. Voorhees. Query Expansion Using Lexical-Semantic Relations. In *Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, Dublin, Ireland, July 1994.

[WBHW97] 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 on IR research*, pages 126–144, Aberdeen, Scotland, April 1997.

[WBHW98] B.C.M. Wondergem, P. van Bommel, T.W.C. Huibers, and Th.P. van der Weide. Agents in Cyberspace – Towards a Framework for Multi-Agent Systems in Information Discovery. In *Proceedings of the 20th BCS Colloquium on Information Retrieval, IRSG98*, Grenoble, France, 1998.

[WJ95] M. Wooldridge and N.R. Jennings. Intelligent Agents: Theory and Practice. *Knowledge Engineering Review*, 10(2):115–152, 1995.

[WSA+96] B.C.M. Wondergem, J. Simons, A.T. Arampatzis, J. Mackowiak, D. Tarenskeen, and T.W.C. Huibers. Profile Information Filtering Project – Overall Project Plan. Version 0.01, Computing Science Institute, University of Nijmegen, Nijmegen, The Netherlands, 1996.

[Zwa96] V. Zwass. Electronic Commerce: Structures and Issues. In *International Journal of Electronic Commerce*, volume 1, pages 3–23, Fall 1996.