Literature Review: Information Filtering for Knowledge Management

By Nikolaos Nanas
Research Student at the Knowledge Media Institute

Supervised by:
John Domingue
Stuart Watt
Enrico Motta

The Open University
June-July 2001

TABLE OF CONTENTS

1. INTRODUCTION

It is already realized that we have entered the knowledge era. A time when the economic value of knowledge has become greater than the value of physical products. It is not accidental that the stock market value of a number of companies far exceeds the visible assets of their balance sheet (Sveiby, K. E. 1997). This difference accounts for a company's "Intellectual Capital" or more specifically its "Knowledge Assets" (i.e. everything the enterprise knows). in an economy characterized by global competitiveness and constantly shifting markets, these knowledge assets can provide today's companies with the competitive advantage they are looking for. After the successes and failures of previous managerial trends like Total Quality Management (TQM) and Business Process Reengineering (BPR), managers are now realizing that the last untapped resource is the knowledge of the employees and of the organization as a whole. As cited by (Nonaka, I. and Takeuchi, H. 1995), Drucker argues in his latest book that in the new economy, knowledge is not just another resource alongside the traditional factors of production – labor, capital and land – but the only meaningful resource today. As a result, Knowledge Management (KM), i.e. the combination of management principles and technology that seeks to improve the performance of individuals and organizations by maintaining and leveraging the value of knowledge assets, has emerged into a managerial megatrend.

Only in the UK it is predicted that the market for knowledge management software will grow from $515 million in 1999 to $3.5 billion in 2004 (Ovum 1999). At the same period the knowledge management services market is expected to grow from $2.6 billion to $8.8 billion. This extreme growth in the knowledge management market has of course its drawbacks. KM is avidly championed by numerous technology vendors that "claim" to provide a wide range of products and services. Without doubt the above abstract definition of knowledge management allows for its instantiation in many different and some times conflicting ways. As with any discipline that lacks a recognized unifying paradigm, various approaches to knowledge management has emerged. This document aims at reviewing some of the approaches to knowledge management, mainly from a technological perspective. Of course it is a common belief that Knowledge Management technology that does not take into account the human political and cultural issues associated with the adoption of any knowledge management system is bound to failure.

The following section establishes the foundations of Knowledge Management. We formulate our perception of the concepts of "knowledge" and "information" and of their interrelation. In Section 3 we present an overview of existing approaches to knowledge management and identify the requirements of a successful KM system. These requirements point towards the direction of Information Filtering. The user of Information Filtering technology for the development of KM applications is an emerging trend. The domain however of Information Filtering is very broad. To establish an understanding of what Information Filtering stands for we investigate in Section 4 its relations with Information Retrieval and Text Categorization. Finally Section 5 presents a number of IF systems, distinguishing at the same time between systems that have the ability to learn, adapt, and evolve.

2. KNOWLEDGE MANAGEMENT FOUNDATIONS

2.1. The Japanese point of view.

In 1995 two Japanese academics, Ikujiro Nonaka and Hirotaka Takeuchi, published the *Knowledge-Creating Company* (Nonaka, I. and Takeuchi, H. 1995), a groundbreaking study of knowledge generation and use in Japanese firms. Nonaka and Takeuchi argue that the traditional western view of organizations (inherited from Cartesian dualism) as a mechanism that processes external information in order to adapt to new circumstances, does not explain innovation. Instead they propose a theory of *Organizational Knowledge Creation*, which they describe as *"the capability of a company as a whole to create new knowledge, disseminate it through the organization, and embody it in products, services, and systems"*.

More specifically, Nonaka and Takeuchi draw on Polyani's distinction between *tacit knowledge* and *explicit knowledge*. A distinction that has become the cornerstone of most theories and frameworks for KM. Tacit knowledge is personal, context-specific, and therefore hard to formalize and communicate. It is highly ingrained into action. It is the knowledge that although allows us to ride a bicycle, we find it difficult to articulate it. Explicit knowledge on the other hand, is knowledge that we can capture and communicate in terms of reports, articles, manuals, blueprints etc. Tacit and exmplicit knowledge account for one of the dimensions of a two-dimensional knowledge creating space. The second dimension of this space comprises the levels of knowledge creating entities (individual, group, organizational and inter-organizational).

Nonaka's and Takeuchi's basic claim is that knowledge creation takes place in this two-dimensional space *"through the social interaction between tacit and explicit knowledge"* and takes the form of a spiral that starts at the individual level and expands to larger communities of interaction (higher level entities). They distinguish four modes of interaction or *"knowledge conversion"*, between tacit and explicit knowledge (fig. 1):

Socialization (from tacit to tacit) creates new tacit knowledge through the sharing of experiences. Socialization describes the kind of learning performed by an apprentice when he observes his master in order to acquire his skills and technical know-how. It is learning by sharing experience. The mere transfer of information will often make little sense, if it is abstracted from associated emotions and specific context in which shared experiences are embedded.

Externalization (from tacit to explicit) creates new explicit knowledge by expressing tacit knowledge in terms of more explicit structures like metaphors, analogies, concepts, hypotheses or models. Nonaka and Takeuchi state that this is the most critical of the four modes of knowledge conversion in terms of knowledge creation. New explicit knowledge is created and can then be communicated.

Combination (from explicit to explicit) creates new explicit knowledge by integrating explicit knowledge entities into larger and more expressive knowledge systems. Individuals exchange and combine explicit knowledge when communicating and especially when collaborating. The importance of collaboration for knowledge management is also studied by Clarke and Cooper in (Clarke, P. and Cooper, M. 2000).

"*Internalization (from explicit to tacit) is the process of embodying explicit knowledge into tacit knowledge"*. This mode of knowledge conversation is closely related to "learning by doing" and is facilitated when the explicit knowledge is diagrammed or verbalized into documents, manuals, or oral stories.

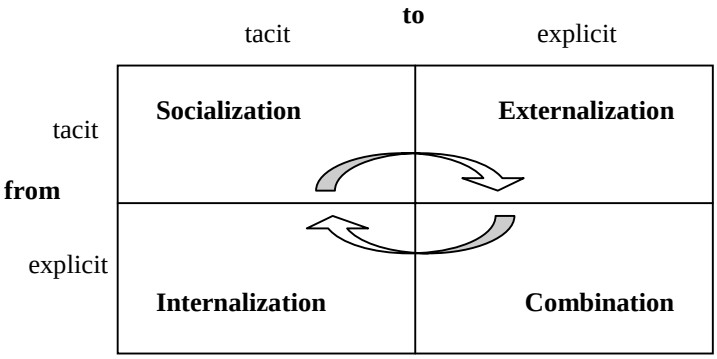

Figure 1.
Four modes of knowledge conversion

Of course none of the above knowledge conversion modes can sustain knowledge creation by itself. As already mentioned Nonaka's and Takeuchi's theory attributes organizational knowledge creation to the interaction between tacit and explicit knowledge through the interweaving of the above knowledge conversion modes. The result is an expanding spiral that starts with the creation of new tacit knowledge by socialization, its externalization to explicit knowledge, its combination with other explicit knowledge and finally back to internalization as individual tacit knowledge. Evidently, tacit knowledge of individuals is the basis of knowledge creation. This is where the spiral starts. An organization cannot create knowledge by itself. However, although Nonaka and Takeuchi acknowledge the importance of the tacit knowledge of the individuals, they focus their study at higher organizational levels, based on feedback provided by middle and higher level managers. Instead, we focus our research at the individual level. Our goal is to investigate the ways individuals can be supported in externalizing their tacit knowledge and creating new tacit knowledge through internalization. What is the role technology can play in amplifying these processes? What has been done so far and what can be done? The value of Nonaka's and Takeuchi's study is of course indisputable and is acknowledged by its broad acceptance. It will be used for a better understanding of the different approaches to knowledge management by identifying the knowledge conversion process they are focusing at. We will avoid committing ourselves to concrete characterization frameworks like the one proposed by (Newman, B. D. and Conrad, K. W. 2000). We prefer taking a discerning stance towards the different approaches to knowledge management than a sterile characterization one.

2.2. A more clarifying point of view: The role of action.

Before however, moving into presenting some of the approaches to knowledge management, it is pertinent to review the processes of internalization and externalization of knowledge from a different point of view. This is the view presented by Scott Cook and John Seely Brown in their seminal paper (Cook, S. D. N. and Seely Brown, J. 1999). Cook and Brown debate that knowledge can be "converted". They argue that *"explicit, tacit, individual and group knowledge are four distinct and coequal forms of knowledge, each one able of performing work the other cannot"*. Consequently, none of these types of knowledge can be derived from or converted into another and there is no reason in assigning greater importance to one of them. Cook and Brown also suggest extending the traditional *"epistemology of possession"*, which treats knowledge as something that can be possessed, by adding a parallel *"epistemology of practice"*. More specifically, they contend that not all of what is known is possessed. Some of it is part of human action itself. It is what they call *"knowing"*. Based on these contentions they finally state that *"new knowledge and knowing lies in the use of knowledge as a tool of knowing within situated interaction with the social and physical world"* (fig. 2). It is thus this *generative dance* that is the key to knowledge creation.

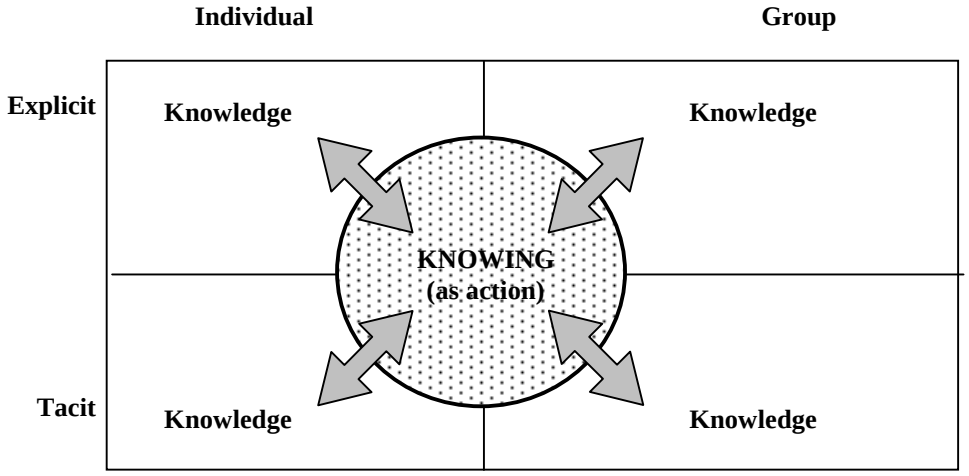

Figure 2. *Knowledge and Knowing. Bridging Epistemologies.*

Cook and Brown also adopt Polyani's distinction between tacit and explicit knowledge. They use the bicycle-riding example to clarify their differences. Thus tacit knowledge is what is known (possessed) by someone that is able to ride a bicycle. We would use explicit knowledge to describe to someone that does not know how to ride a bicycle, how he should move in order to stay upright. However, whatever the amount of explicit knowledge we provide him with, he will never be able to ride a bicycle unless he actually gets involved with the action of riding a bicycle. Explicit knowledge can only aid him to acquire (internalize) the tacit knowledge he needs in order to be able to ride. On the other hand, one can use the riding action to formulate (externalize) the explicit knowledge he needs in order to describe to someone else how he should ride a bicycle.

Cook and Brown use the above example to come up with the following conclusion about the relations and differences between tacit and explicit knowledge:

- Each form of knowledge (tacit, explicit) can facilitate the acquisition of the other.
- Each form of knowledge cannot be used by itself to acquire the other. Situated action must also take place.
- Facilitating the acquisition does not mean conversion from one form of knowledge to the other. None of these types of knowledge can be turned into the other.

Interestingly Cook and Brown avoid the use of the terms *"internalization"* and *"externalization"* since these terms do not comply semantically with the last of the above conclusions. Both terms imply a change of state. Instead, they present each one of the forms of knowledge, as a catalyst to the *"acquisition"* of the other in the context of action. Although we agree with their view for the reasons presented later on, we will never the less use the terms *"internalization"* and *"externalization"* in the rest of this review, without however implying any conversion between types of knowledge.

2.3. Converting and Connecting.

We conclude our review of the foundational literature on knowledge management by presenting O'Leary's seminal paper (O'Leary, D. E. 1998). Based on his study on the knowledge management practices of big consulting companies O'Leary defines knowledge management as *"the process of converting knowledge from the sources accessible to an organization and connecting people with that knowledge"*. He expands upon this definition by presenting in detail a number of converting and connecting processes that he argues to be critical to knowledge management. More specifically he decomposes knowledge management in the following process:

Converting …

… *individual to group-available knowledge.* Individual knowledge has to be made accessible by the group. This implies the identification of knowledge that is desirable to share and its collection in a form that can be generated and reused.

… *data to knowledge.* O'Leary points to the importance of knowledge discovery as part of a knowledge management initiative. He borrows Piatensky-Shapiro's and Frawley's definition of knowledge discovery as *"the nontrivial extraction of implicit, previously unkown, and potentially useful information from data"*.

… *text to knowledge.* Textual information like news articles is identified as a potential source of knowledge. Intelligent agents can be used to facilitate users in generating knowledge from this kind of textual information.

Connecting …

… *people to knowledge.* Knowledge creation and identification can result in huge knowledge repositories that are difficult to navigate. Specialized tools can be used to help individuals understand what is available or provide them with personalized services (e.g. InfoFinder (Krulwich, B. and Burkey, C. 1997)).

… *knowledge to knowledge.* Although documents have been the basis of most knowledge management systems, documents by themselves are of little value if they are not appropriately linked together. This can be done using hypertext, preferably in a way that combines both a vertical, structured model ("Vatican" model) and a horizontal, flat model ("World Wide Web" model). Whatever the model however, hypertext should facilitate personalized views, since users do not look for the same information in the same way. Multiple routes to the same destination have to be provided.

… *people to people.* Communication networks allow individuals to share what they know. This can be a very important asset because people are the greatest source of knowledge. Thus people should be made able to advertise their skills while technologies like intelligent agents can be used to identify experts that could prove useful.

… *knowledge to people.* Looking for useful knowledge is neither an effortless nor a timeless process. So letting the users look for the information they need ("pull" strategy) can result in unfound and unused knowledge. Alternatives technologies that "push" knowledge to the user have to be investigated.

O'Leary's paper instances the above higher level theories by reciting specific services that knowledge management technologies should support. Without doubt his categorization of knowledge management technologies complies with Nonaka's modes of knowledge conversion (e.g. people to people complies with Nonaka's socialization). He however brings us closer to a technological perspective to knowledge management with more concrete technological examples. It is also worth noting that he identifies data and information as potential sources of knowledge, although he does not discuss their actual relation. It is one of our aims in this review to investigate the role of information to knowledge management and we thus find it necessary to provide some insight to the interrelated and misunderstood concepts of data, information and knowledge.

2.4. Data, Information and Knowledge.

So far we have learned to distinguish between tacit and explicit knowledge, we have also discussed the role of knowing as part of action in creating new tacit (explicit) knowledge with the aid of explicit (tacit) knowledge, and finally we have identified but not justified, that information is a significant part of knowledge management. To support our focus on information as a critical resource for the creation of new knowledge, we will now try to illuminate the concepts of data, information, and knowledge and clarify their interrelation. The conclusions derived from the following discussion will shape the direction of the rest of this review.

2.4.1. Data and Information.

Davenport and Prusak define data as *"a set of discrete, objective facts about events"* (Davenport, T. H. and Prusak, L. 1998). Data does not have meaning of itself. It describes part of what happened without relation to other things. As a result, data provide no justifications or interpretations. According to Davenport's and Prusak's citation of Peter Drucker *"information is data endowed with relevance and purpose"*. In other words information can be acquired from data by giving them meaning in terms of relational connections (Bellinger, G., et al.). So data can be seen as the raw material for the creation of information. However, although information provides the meaning inherent in the relations between data, it cannot provide an explanation of "why" the data is what it is (Bellinger, G., et al.).

2.4.2. Information and Knowledge.

Information is a means for communicating a message that is able of changing the receiver's perception of a situation and affect his judgements and decisions (Davenport and Prusak, 1998). Information can provide the receiver with a new way of interpreting objects and events by highlighting unexpected connections and implying unconsidered constraints (Nonaka, I. and Takeuchi, H. 1995). What both Nonaka and Davenport imply is that information is a necessary material that aids in eliciting and constructing knowledge. More specifically and according to Nonaka's citation of Dretske, *"information is commodity capable of yielding knowledge, and what information a signal carries is what we can learn from it. … Knowledge is identified with information-produced (or sustained) belief"*. Davenport and Prusak complement this argument by providing four human related activities that can result in the creation of knowledge based on the information received. These activities are:

Comparison: how does the information received under this context compare to previous situations.

Consequences: what are the implications (e.g. constraints or new insights) the received information has on the decision process of the current situation (context).

Connections: what are the relations of the received information to other bits of information.

Conversation: what do other people think about this information.

It is evident by the above activities, that information's perceived meaning and thus the knowledge that can be created based on it, is dependent on context and that the receiver (his beliefs and commitment), his situation and his social interaction define this context (Nonaka, I. and Takeuchi, H. 1995). So the same piece of information can "give birth" to different knowledge, according to the context of the receiver. At the same time it is within the human receiver and his interaction with other humans that the above activities take place.

2.4.3. Knowledge, Action and … Information.

We can now come to the conclusion that knowledge does not only allow humans to act (e.g. ride a bicycle) but also to perceive information and adjust their actions appropriately according to what they have perceived. Hence we would like to adopt Davenport's and Prusak's definition of knowledge as:

… a fluid mix of framed experience, values, contextual information, and expert insight that provides a framework for evaluating and incorporating new experiences and information. It originates and is applied in the minds of knowers. In organizations, it often becomes embedded in documents or repositories but also in organizational routines, processes, practices, and norms.

Therefore, the most value-adding characteristic of knowledge is its ability to create new knowledge by perceiving and evaluating information in the context of action. It is now made clear why we have already agreed with the great importance Cook and Brown have assigned to knowing as part of action. Action implies decision and defines context. However decision making in any knowledge intensive task involves ambiguity and uncertainty. As already mentioned, information has the potential of providing the extra insights and constraints needed by the decision maker in order to reduce uncertainty and decide on a specific action. New knowledge is created as a side effect of the decision process itself. It is not extracted directly from the information used in support of the decision process. Borghoff's and Pareschi's definition of knowledge work comes in support of this discussion on the importance of information to the creation of new knowledge (Borghoff, U. M. and Pareschi, R. 1998).

Knowledge Work: a new type of intellectual work that is all about making sense, namely, about turning information into knowledge through the interpretation of incoming highly non-standardized information for the purposes of problem solving and decision making.

To continue on the characteristics of knowledge, we should also add that knowledge is complex and adaptive (Bellinger, G., et al.). Knowledge has the ability to deal with complex situations, situations that don't fit to what is already known, in a complex way. As a result, and in light of what is already known, it judges and refines itself in response to new situations and information (Davenport, T. H. and Prusak, L. 1998). Once more information plays a catalytic role. Of course all these characteristics of knowledge add up to what makes knowledge the most valuable asset in the current organizational endeavor for competitive advantage, i.e. its ability to increase with use.

2.4.4. Internalization and Externalization revisited: putting them all together.

Now that we have at least tried to clarify the concepts of data, information and knowledge and their interrelations, it is a good idea to revisit the concepts of tacit and explicit knowledge and of their creation through internalization and externalization. We adopt at this point the definition of knowledge by (Davies, N. J., et al. 1998) to redefine explicit knowledge as "*information transformed into a capability for effective action*",

which implicitly results in the creation of new knowledge. Of course this statement can raise a lot of arguments and it is not our intention at this point of our ongoing research to get involved in such philosophical juxtapositions. The question "Can knowledge be communicated?" is a research question in each own right, but not the research question we are trying to address. It is however our duty to raise questions like: "Can we call knowledge something that does not have the ability to create new knowledge?", "Can we call knowledge something that can not judge and refine itself?", and finally "If knowledge can be communicated then why does learning involves such a long process of trial and error? Why can't we just give to someone the knowledge of riding a bicycle?".

Whatever the answers to the above questions, it is interesting to note that this understanding of explicit knowledge provides a unified view of the theories we have reviewed so far. First of all claims like,"*tacit knowledge is acquired on its own; it is not made out of explicit knowledge*" (Cook, S. D. N. and Seely Brown, J. 1999), can be clarified by just replacing the term "*explicit knowledge*" with "*information*". Furthermore the ambiguity in the use of the terms "knowledge" and "information" in expressions like "*Because these knowledge bases are limited to a single type of information…*" (O'Leary, D. E. 1998) that is characteristic in the knowledge management literature, is now understood. Finally, we can also redefine the concepts of internalization and externalization as:

Internalization is the process of acquiring new knowledge through action that is aided by enlightening information.

Externalization is the process of producing new information, that has the inherent capacity to be enlightening, as the result of action.

In other words, if we provide a knowledge worker, i.e. the intellectual individual working on a knowledge intensive task (knowledge work), with the appropriate information that can make visible previously invisible meanings or can shed light on unexpected connections (whence the characterization enlightening), then the individual can be aided to reflect on his decision process. New tacit knowledge is then acquired as a result of that same process. In addition and during the process itself the individual produces new information that at least implicitly reflects part of his decision process. Thus this information can prove enlightening under similar circumstances in the future.

The importance we have assigned to information as a critical resource for the creation of new knowledge is also reflected to most of the approaches to knowledge. As we will see in the next section most approaches to knowledge management are in fact trying to solve the problem of providing the individual with the right information at the right time given his context. They try to find out what exactly he needs by identifying what exactly he is doing or what exactly the information is about or both. We will discuss these approaches and highlight their advantages and disadvantages. The importance of information for knowledge management also pinpoints towards the direction of information filtering. Later in this review we will investigate this research area and we will see that information filtering systems do not only have the inherent ability to deal with information access but have also the potential to overcome some of the disadvantages of existing approaches to knowledge management.

3. APPROACHES TO KNOWLEDGE MANAGEMENT

In the endeavor for the development of successful knowledge management (KM) applications, Artificial Intelligence (AI) plays one very significant if not the most significant role. For researchers in classical AI, that has dominated the domain for the last 30 years, knowledge management sounded by definition as a new application domain for knowledge bases, knowledge acquisition, knowledge representation and ontologies. However the transition has proven to be not as smooth as it literally sounds. The following discussion will reveal the limitations of knowledge based approaches to knowledge and at the same time it will define their niche in the domain based on their strengths.

3.1. Knowledge based approaches to knowledge management.

An overview of the role knowledge bases and ontologies can play in the development of knowledge management systems is presented by (O'Leary, D. E. 1998). O'Leary argues that knowledge bases and especially best practices knowledge bases are a necessary prerequisite for successful KM applications: *"… the existence of best-practices knowledge bases signals the extent of development of a KM system: less developed KM systems generally do not have best-practices databases; more developed systems do"*. Based on a study of KM practices in consulting firms he distinguishes between different kinds of knowledge bases and then describes the different issues that have to be taken into account when developing a knowledge base given the difficulty of the task. Then he specializes his presentation in best-practices knowledge bases and in particular examples in three large consulting companies.

Having stressed the importance of knowledge bases, he suggests the use of ontologies for their more effective use. He defines ontologies in the context of knowledge management, as *"specifications of discourse in the form of a shared vocabulary"*, and distinguishes between five applications area where ontologies can prove useful. These areas are:

- The definition of the scope of discussion groups so that users can know where to raise a specific issue.
- The provision of search capabilities by determining the topics residing in a knowledge base.
- Filtering capabilities can be provided based on the underlying ontology (or ontologies). Users can use the ontology to specify keywords or concepts that capture the nature of the desired knowledge.
- Reusing artifacts archived across the common dimensions of an ontology. This way the similarity of an artifact to the current situation can be determined.
- Finally ontologies can be used to facilitate collaboration either by defining a common language or by facilitating the identification of the appropriate expert.

O’Leary concludes with a presentation of the desirable characteristics an ontology should have and the tools that can help the development of ontologies and especially their integration with knowledge bases.

It is obvious that according to O’Leary ontologies can provide the required means for realizing all of the *converting* and *connecting* processes he describes in (O'Leary, D. E. 1998). We will however concentrate on the processes, which are directly related to the individual, namely, connecting people to knowledge (information retrieval) and connecting knowledge to people (information filtering). A common solution to these problems is to annotate each information entity (e.g. document) using an ontology. After being annotated each entity can be retrieved using the same ontology and intelligence inference on it. The following examples will provide a clearer understanding of this approach.

3.1.1. Annotating HTML pages.

Annotation for example of html documents based on an ontology, can be achieved using an extension to the HTML mark up language (Benjamins, R. V. 1998). Meta-data, like the author of the page or its topic, can then be assigned to the page using the extra attributes provided by the extension. This results in a distributed approach where each individual annotates each own documents (e.g. his home page). There is not central repository of documents. The annotated documents are retrieved with the help of an ontology based brokering service called Ontobroker, which consists of a web crawler (called Ontocrawler), an inference engine and a query interface. The Ontocrawler is responsible for searching through the pages and collecting their annotations. The Ontobroker then translates the collected annotation into facts expressed using the underlying representation language. Users can now use an interface to appropriately form queries that are received by the inference engine which responsible for finding relevant pages based on the formulated facts.

The described approach is being tested using the Knowledge Annotation Initiative of the Knowledge Acquisition Community as a test case. Seven ontologies form the underlying ontological substrate. The researchers involved in the community are required to annotate their home pages or any other html document they consider interesting and then register them to the Ontocrawler. Queries can be formulated by initially browsing the ontology using a hyperbolic visualization tool and then by filling the form that appears when clicking on one of the ontological concepts.

Benjamins acknowledges some of the risks of the proposed solution, like its dependence on the researchers’ willingness to contribute (given the difficulty of the annotation process), its ability to scale up to an environment involving thousand of pages and most importantly the effect that changes to the underlying ontology can have to the annotations on the pages. He however supports the approach through a brief comparison with keyword based approaches. He argues that keyword based approaches suffer by the large number of results they return to the user and their inability to present the results in a coherent way. Moreover the ontology-based approach has the ability of accessing implicit knowledge which is something keyword-based approaches cannot do.

3.1.2. Web publishing with Planet-Onto

Annotation of information entities and more specifically news stories is also supported by the *Planet-Onto* architecture presented by (Domingue, J. and Motta, E. 2000). Planet-Onto provides an integrated set of tools that supports and augments the publish/find/read processes related to a central news server called *KMI-Planet.* More specifically, journalists (i.e. users that have the right to publish stories) can submit a story by just sending an email to the KMI-Planet server. The journalist submitting the story or a knowledge engineer can annotate the story using the underlying ontology. The latest defines the concepts needed to describe events related to academic life, given that the system is currently used in an academic environment. This process is supported by the *Knote* tool, which provides the user with a form-based interface. The user can use this interface to classify the event described in the story in terms of the event types defined by the ontology and also create a new instance of the event type he has selected to assign the story to.

Users can of course browse through the stories stored in KMI-Planet in a traditional way, but the annotation of stories in terms of the underlying ontology allows for the development of search and filtering services. Users can search for relevant stories using the *Lois* interface to construct queries, which are formulated as a conjunction of ontological concepts. Relevant stories can then be found using deductive knowledge retrieval. Domingue and Motta however identify the importance of pushing stories to readers instead of waiting for them to search for them. This has resulted in the development of the *NewsBoy* agent that provides readers with personalized alerts on potentially interesting stories. Each reader can specify a number of queries using Lois and thus construct a personal profile. New annotated stories are matched against the user profiles and interested users are appropriately notified.

3.1.3. Proactive knowledge delivery

This latest fact, that it is more appropriate for a KM system to act proactively by pushing information to the user, is further emphasized by (Abecker, A., et al. 1999, Abecker, A., et al. 2000). Abecker argues that is usually the case that users are not aware of the existence of useful information and even if they do, they do not know where and how to look for it. Furthermore, looking for useful information is not a timeless and effortless process and therefore individuals are not always willing to stop their ongoing work to do so. Instead, Abecker proposes the use of an active, context-sensitive KM system (*KnowMore*), that proactively pushes information to the individual according to his task at hand (i.e. his context). More specifically, each knowledge intensive task (KIT) is characterized by a number of generic queries that are instantiated on the fly during workflow enactment with the help of a workflow engine. Both the variables forming the generic queries and the information sources are described in terms of a domain ontology. Thus, as in the case of Planet-Onto intelligent knowledge retrieval can be performed to retrieve the information that best match the instantiated queries. Furthermore, information created during a KIT is appropriately indexed given the workflow description of the KIT for future retrieval. Describing the content of the information entities in terms of an ontology also achieves integration of heterogeneous information sources.

3.2. Advantages and disadvantages of knowledge-based KM systems.

It is made obvious by the above discussion of knowledge-based approaches to KM that the overall goal of these systems is to provide the individual with the information *relevant* to his needs or interests. This is achieved by first annotating or describing the information entities in terms of an underlying ontology. Deductive algorithms and heuristics can then be applied on the ontology to define the relevance of the annotated information entities to the individual's needs or interests, that are also expressed in terms of the same ontology. Ontologies therefore provide a common machine-readable language for comprehensively and consistently interrelating information entities to other information entities and to user queries. The algorithms used not only return unambiguous results but can also present the rational for a given answer. Finally, the use of meta-information to describe the content of an information entity allows for the retrieval of non textual information like images and blueprints.

However, knowledge-based approaches suffer from a number of disadvantages that hinder their successful application to the domain of knowledge management. As the following discussion will reveal, the problems related to the use of knowledge-based systems for KM are inherent in the use of formalized abstractions like the ones expressed by an ontology. These problems can be distinguished between problems in the ways relevance is defined and problems of user acceptance.

3.2.1. Relevance in four dimensions.

To formulate our further discussion we use the four dimensional representation of relevance by (Mizzaro, S. 1998) (fig. 3). According to Mizzaro relevance can be defined in a four dimensional space where each one of the dimensions represent:

Information Recourses (InfRes={Surrogate, Document, Information}.

- *Document,* the physical entity storing information.
- *Surrogate,* a meta-level representation of the document.
- *Information,* the actual information stored in the document.

Representation of the user's problem (Repr = {Query, Request, PIN, RIN}).

- *Real Information Need (RIN),* the actual information that the user needs or is interested in.
- *Perceived Information Need (PIN),* the user perceives his RIN and creates a mental representation of it. Usually the user does not know exactly what he needs.
- *Request,* the user expresses the PIN in "human" language.
- *Query,* the user translates his request to a "system" language.

Time

The above two dimensions define relevance as a point in a two dimensional space. Relevance can thus be seen as the relation between two entities, one from each one of the above two groups. However, whatever the type of relevance (e.g. surrogate-query), relevance is always time dependent. Something that is relevant now can be irrelevant after some time and vice versa. The individual usually reflects on his actions and the information he has received so far until he finally perceives more clearly his RIN at t(f). Thus relevance evolves over time during the execution of a knowledge intensive task. KM systems have to be able to adjust to these changes of relevance over time.

Components (Comp = P(topic, task, context) - ∅ =
{{topic}, {task}, {context},
{topic, task}, {topic, context}, {task, context},
{topic, task, context}}).

- *Topic,* the interests and/or the domain of expertise of the user.
- *Task,* the knowledge intensive task that the user is trying to accomplish.
- *Context*, everything not pertaining to topic and task, but however affects the relevance of information (e.g. social interaction)

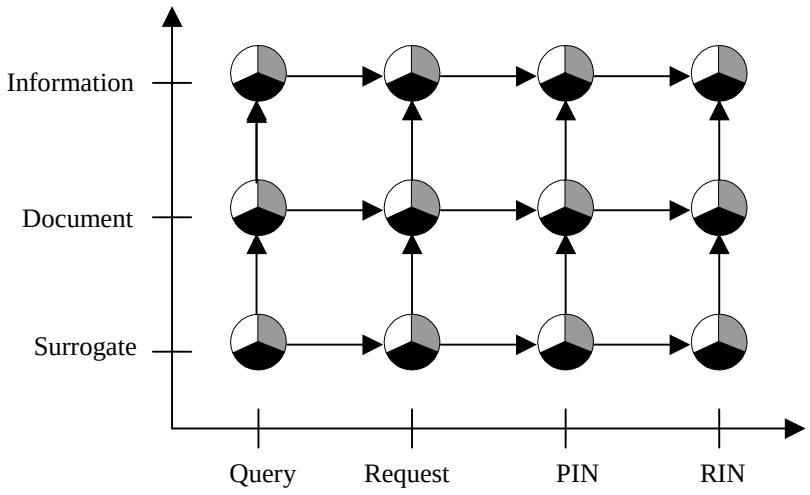

Figure 3. Relevance in three dimensions. (*time* is the fourth not depicted dimension)

We can now define relevance as a point in the above four-dimensional space. Obviously the optimum relevance that can be achieved is of type *RelO={information, RIN, t(f), {topic, task, context}.* Information of such relevance has the greatest potential of being useful (enlightening) to the receiver. Of course it is extremely difficult if not impossible to achieve optimum relevance of the provided information. Most knowledge-based systems for instance achieve relevance of type *Rel={surrogate, query, t, topic}.* More specifically:

(Request → Query) Users find it difficult to express their request in terms of a formal language. Formal representations are not always comprehensive by users who are not knowledge engineers (Buckingham Shum, S. 1998). Even when this representations are hidden behind an interface, like the hyperbolic and form based interfaces used by the presented approaches, users do not fully understand their declarations in a structured language (Shipman, F. M. and Marshall, C. C. 1999, Shipman, F. M. and McCall, R. 1994). In addition the use of an interface for the generation of a formally structured query adds an extra burden to the already busy employees. Users have to take many extra steps and make additional decisions to specify their request in terms of a structured query (Shipman, F. M. and Marshall, C. C. 1999). This can also hinder the easy adoption of a knowledge management system as sited by (Masterton, S. and Watt, S. 2000).

(Document → Surrogate) Knowledge-based approaches assess relevance based on meta-level information about the document (information entity) and not based on the document's actual content. First of all this implies that only documents that have been already annotated with meta-level information can be retrieved. Thus dynamic and potentially useful sources of information like the internet and email transactions are excluded by default. It would be physically impossible for any individual to be manually annotating each and every piece of interesting or useful information he comes across during his every day work. But even if we could annotate any accessible piece of information, using for example automatic information extraction techniques, this would only partially solve the problem. (Buckingham Shum, S. 1998) states that ontologies express a simplification of the world, maybe the most important part of it, that inevitably factors out certain classes of information simply because they are hard to formalize. As a result any piece of information can only be classified according to this restricted view of the world. Thus, there is usually a mismatch between the user's understanding of the information and the choices that a formal scheme provides him with in order to represent it (Shipman, F. M. and Marshall, C. C. 1999). This problem is further augmented by the vocabulary problem, i.e. the use of different terms by different people to express the same topic (Furnas, G. W., et al. 1987). Even in the simple case of the classification scheme provided by the UseNet news server, studies reveal an inconsistency in the way users classify the same news stories.

(Information → Document) We believe that it is impossible to know in advance which bits of the information stored in a document is of interest or needed by the user. Hence we agree with Frederick Brooks' cooperation formula as cited by (Masterton, S. and Watt, S. 2000). According to Brooks IA>AI, which means that the combination of intelligent assistance with the user (Intelligent-Amplification) is much more powerful than any separate AI system. In other words a system should provide the user with information that it "believes" is relevant, but it is the user who decides on the usefulness of the received information. If for example a system presents information to the user in order of decreasing relevance there is always the possibility that it will be the X^{th} document that will prove most useful and not the

first one. There is no way to deterministically assess the usefulness or interestingness of the information stored in a document in advance.

(Time) Formalized representation are not usually flexible enough (Buckingham Shum, S. 1998). Most of the time costly maintenance has to be performed by knowledge engineers or by the users themselves to adapt the representation to changes in understanding and in the domain being represented. It is however extremely difficult to represent the constantly shifting and task dependent user interests and needs. A flexible, adjustable solution is thus required.

(Topic) The *Topic* component can be modeled using a domain ontology. However, as already mentioned ontologies express a partial view of the world. As a result a problem arises when we try to express the user interests and needs in a user profile using the concepts defined by the ontology. Not only it is difficult for the user to express his interests in terms of the formalized concepts provided, as mentioned earlier, but also the diversity of the various profiles is restricted to the classification choices provided by the system. We argue that each user is a class of his own and ontologies cannot cope with such diversity. Hence, instead of a top down approach that tries to express user profiles in terms of predefined concepts, a more flexible bottom up approach that adjustably (see previous paragraph) builds a distinct user profile for each one of the involved individuals is needed.

(Task) A task can be modeled using a process model instantiated by a workflow engine as in the case of the KnowMore system discussed above (Abecker, A., et al. 1999, Abecker, A., et al. 2000). It is however a common fact that workers seldom follow strict procedures. The procedures followed are usually exceptions to the prescribed form (Shipman, F. M. and Marshall, C. C. 1999). As a result systems based on standard procedures can be too brittle and bring processes to a halt (Grudin, J. 1994). More flexible solutions that can tap into the existing formal and informal ways of doing things are needed.

(Context) Context is the most difficult component to deal with. An individual's beliefs, intentions and social interactions are too difficult to identify and of course to be modeled. They are however implicitly reflected in the way the individual perceives his information need and thus evaluates the usefulness or interestingness of the received information. If a system is able to learn from and adjust to the way a user evaluates the relevance of information then, at least to some extent, these contextual parameters will be reflected to the way the system evaluates the relevance of information.

The above discussion has revealed the limitations of knowledge-based systems to support the individual working on a knowledge intensive task with relevant enough information and the difficulties users face in interacting with such systems which hinders their easy adoption. Of course this does not exclude the use of knowledge-based systems in the knowledge management domain. We will however agree with both Buckingham Shum and Shipman who state:

"KM technologies should formalize only knowledge which is stable and sanctioned" (Buckingham Shum, S. 1998)

" Such specialized formal representations are possible for well-defined tasks, but general tasks like analysis and design evolve over time and vary from person to person." (Shipman, F. M. and Marshall, C. C. 1999)

3.3. Other approaches to Knowledge Management.

Of course knowledge-based systems is not the only way of providing knowledge management services. It is however the dominant way especially for systems that try to target the individual knowledge worker and his everyday work practices. Most alternative solutions provide specialized services that only facilitate part of the processes composing a KM infrastructure. Argumentation tools based on hypertext for example are used for structuring discussions during meetings (Buckingham Shum, S. 1998, Selvin, A. M. 1999, Shipman, F. M. and McCall, R. J. 1997); Conklin). These systems enable the capturing of the content of discussions in a structured way and thus facilitate, first of all the progress of the discussion itself, and also the retrieval of such a structured representation of the discussion during future discussions. As a result redundancy in the issues discussed and the agreed solutions is avoided. Hypertext can also play other roles in a KM initiative. Integrating hypertext to existing applications (e.g. design) applications and linking documents to the artifacts produced during a task can add task relevance to the hypertext documents (Reeves, B. and Shipman, F. 1992). We will investigate this use of hypertext as part of our future research. We however believe that the goal of hypertext is not to actively provide the individual with the information he needs. Hypertext systems count more on the *Serendipity Effect*. Users find what they are looking for by navigating the information space in a structured way. Any personalization provided by adaptive hypertext (Brusilovsky, P. 1996), is only in the form of personalized presentations of pages and personalized navigation instructions and not personalized delivery of information.

3.4. Information Filtering for Knowledge Management.

The above discussion on the disadvantages of knowledge-based systems also reveals a number of higher level requirements for a system that provides the user with the information that he needs to accomplish his task and thus acquire new knowledge.

A first requirement is that the system should be able to actively provide the individual with highly relevant information given his needs and interests. This does not mean that the system has to be omniscient or omnipotent. It has just to be able to provide the individual with relevant enough information so that there are increased possibilities that he will find what he needs in the presented information. As already mentioned, in most of the cases even an expert does not know what exactly he needs or even if he knows he does not know how to look for it. It is however more likely that an expert involved in a knowledge intensive task will recognize the usefulness of the information he receives. Active support also means that the individual's interaction with the system is

minimized. The user does not have to express his needs in any formal representation. Minimized interaction means easier system adoption by the users.

Furthermore, such a system should be flexible enough to adjust to changes in the needs and interests of the user. This implies that the system should be able to learn from its failures and successes. A side effect of this learning capability is that the user's needs and interests are eventually reflected in the system's underlying mechanisms. Thus although the user does not have to explicitly declare his topic or task these components of relevance are implicitly incorporated by the system and reflected in the documents it "believes" to be of relevance to the user.

Evidently, the above requirements point towards the direction of personalized information filtering. This established discipline, that has only recently become fashionable with the emergence of intelligent information agents and personal assistants, is now entering the domain of knowledge management as acknowledged by (Borghoff, U. M. and Pareschi, R. 1997). The *Knowledge Pump* system for example, uses community-centered collaborative filtering to disseminate information according to user recommendations and the established "trust" between users within a given category of documents (Glance, N., et al. 1999). According to this approach, a document is presented to a user if it was highly rated by another individual whose judgements the user trusts. For more details on collaborative filtering see (Konstan, J. A., et al. 1997, Shardanand, U. and Maes, P. 1995).

A more interesting however approach is followed by the *Knowledge Sharing Environment (KSE)* system. *KSE is a system of information agents for organizing, summarizing and sharing knowledge from a number of sources, including WWW, an organization's intranet or from other users* (Davies, N. J., et al. 1998). Each user has his own information agent, which maintains a user profile that represents the user's information needs and interests. The user profile comprise a number of user specified phrases or terms and is refined according to its usage as it will be explained bellow. KSE agents provide a number of knowledge management services. More specifically:

- Users can add new information to the system. This information can be a web page annotated with some comments, individual notes by the user themselves or an information entity copied from some other application. When ever the user provides the system with new information this information is matched against his profile and if the match is not good enough the agent suggests to the user new terms and phrases to be added to the profile.
- Information that is added to the system is also matched against the profiles of other users and those users that are interested enough are notified through an e-mail message. A user receiving information can give feedback to the system and as a result his profile is refined appropriately.
- KSE also enables a user to find other users with similar interests. In this case the user's profile is matched against the profiles of all other users and the system presents him with the users that have the best matching profiles.
- Finally, in the same fashion a user can find all users that are interested in a specific document. The document is then matched against the profiles of all other users and those that are interested enough are presented to the user.

The KSE system example exhibits the ability of information filtering technology to support most of the converting and connecting processes described by O'Leary, without the use of any formal representation. Its flexibility and the minimum user effort needed has made the system easily acceptable and it is already used by British Telecommunications. However, the underlying information filtering technology used is quite basic. KSE is using a vector space representation of documents and profiles with a Boolean weighting scheme (see next section). We believe that a most advanced information filtering technology can support even more elaborate and effective KM services. The next section investigates the current state of the art in information filtering from a KM perspective. We will try to identify those characteristics that allow or not the use of existing information filtering technologies in a KM context.

4. INFORMATION FILTERING FOUNDATIONS.

So far we have argued that information is a significant resource for the creation of new knowledge. A knowledge management system should be able to provide the individual working on a knowledge intensive task with relevant information to help him reflect on his decision process and thus acquire new knowledge. We have also seen that in trying to do so, knowledge-based KM systems suffer by certain disadvantages, due to the use of inflexible formal representations of both the information entities and of the user needs and interests. Thus, in order for an information filtering technology to overcome these disadvantages, it should be:

- Able to assess relevance of information entities based on their content and not on some surrogate representation of this content.
- Able to learn from the user and hence minimize his interaction with the system.
- Flexible enough to be able to reflect any changes in the user interests and needs.

These higher level criteria will be used in section 5 for evaluating information filtering technologies in terms of their ability to support the development of a KM system. In this section we will try to clarify what information filtering stands for and its relation with the more traditional disciplines of *information retrieval* and *text categorization*. We will also present some of the algorithms used for *term weighting* which forms a common ground between the three disciplines.

4.1. Information Retrieval, Text Categorization and Information Filtering.

The sudden increase in the availability of digital information especially due to the World Wide Web has brought the problem of information access in sharp focus. Information Retrieval (IR), Text Categorization and Information Filtering (IF) are three disciplines that are trying to cope with this problem, that we have learned to call *"Information Overload"*. As a result of their common goal these three disciplines have some characteristics in common but a lot of differences as well. Therefore, it is pertinent to present IF in relations to IR and Text Categorization.

Traditional IR research focuses on the development of algorithms and models for the retrieval of textual information from document repositories (Manning, C. D. and Schutze, H. 1999). Text Categorization is concerned with the problem of automatically assigning a class label or subject descriptor to texts that belong to the same class (Moens, M. and Dumortier, J. 2000). Categorization of documents facilitates their indexing and thus their subsequent retrieval. Finally, IF is an information access activity that deals with the filtering of a dynamic stream of incoming textual information according to evolving user interests (Foltz, P. W. and Dumais, S. T. 1992).

The differences between IR and IF are analytically presented by (Belkin, N. J. and Croft, W. B. 1992). These differences can be summarized as follows:

- IR systems are concerned with the collection and organization of texts so that users can then easily find a text in the collection. On the other hand, IF is concerned with the removal of textual information from an incoming stream and its dissemination to groups or individuals.
- This incoming stream is usually broadcasted by a dynamic source that produces large amounts of information. On the contrary IR is concerned with the selection of texts from a relatively static repository.
- Filtering is based on descriptions of individual or group interests or needs, that are usually called *profiles*. Retrieval of information is instead based on user specified information needs in the form of a query.
- A query represents a one-time information need while information filtering is concerned with repeated uses of the system by users with long-term, but changing interests and needs.

Information Filtering can also be seen as a special case of text categorization where each user corresponds to two categories, relevant and not relevant documents to the specific user. The basic difference however, is that in contrast to the above categories of IF, which are evolving according to changes to the user interests and needs, categories in Text Categorization are fairly static.

Despite their differences, IR, Text Categorization and IF have a basic similarity. They are mainly concerned with textual information. So at a higher level all three technologies comprise three components: a) a representation of the document, b) a representation of some information class (information need or category) and c) a similarity measure between the previous two. The following three paragraphs discuss each one of these components without however providing any details on specific implementations. The aim of the discussion is to present the general context of information filtering as it is defined by its relation to IR and Text Categorization. In Section 5 we will present some information filtering and user profiling systems and thus different implementations of the above components will be discussed in detail from the point of view of IF.

4.1.1. Document Representation

Usually, documents are not represented using some meta-description. Instead each document can be treated: a) as a collection of letters in the order they appear in the document, b) as the set of unique words that appear in the document and their corresponding frequencies of appearance and c) as the collection of words that appear in the document in the order that they appear. These approaches give rise to increasingly rich representations in terms of semantics. More specifically, in the first case an *n*-gram analysis of the documents is performed, where *n* is a positive integer (Manning, C. D. and Schutze, H. 1999). So for example, a 3-gram analysis treats a text as the set of all triplets of letters that can be constructed by the letters in the document in the order that they appear. This approach results in a representation of the text as a vector in the space of all possible such triplets (Tauritz, D. R., et al. 1999). The dimensionality of such a space increases exponentially with *n* and so most approaches based on this model do not use *n>3*. Of course this representation of a text excludes any semantics and instead represents both the text and the information class in terms of a syntactical space.

In the second case, each document is treated as a *"bag of words"*. The order in which the words appear in the document is not taken into account. If in addition the document is analyzed as part of a collection then it can be represented as a vector in an *n*-dimensional space where *t* is the number of unique words in the collection. This is referred to the *Vector Space Model* and is one of the most widely used models, which is appropriately adopted by all of the discussed technologies. According to the vector space model, absence of a term is indicated by 0 while presence of a term is indicated either by 1 (*binary vector*) or a positive number (*term weight*). A term's weight reflects the discrimination power of the corresponding term. That is, its power to discriminate the document it belongs to from the rest of the documents in the collection. Term weighting is a very significant component of all three technologies and actually stands as a research area of its own. We will discuss term weighting in more detail later on. Now the semantics of such representations that treat a document as a *"bag of words"* correspond to the semantics of the individual words included in the document. Phrases are not taken into account. Furthermore, the vector space model and similar approaches assume that the words in the text are orthogonal and independent (Manning, C. D. and Schutze, H. 1999). This assumption, although not true in most cases, facilitate the use of certain IR models by minimizing the number of parameters that have to be estimated (Losee, R. M. J. 1989).

More semantically rich representations can be used to describe a document if the order of words in the document is taken into account. This way a text is treated as the set of all the phrases it contains or the context of a word is taken into account by considering the words that precede and follow it (Cohen, W. W. and Singer, Y. 1996). Therefore, these approaches take into account the dependence between words, as the importance of a word depends on the phrase it belongs to or on its general context. Additional semantics can be derived if the structure of the documents is known in advance. A document can then be divided into features like the title, the author, the body of the text, e.t.c. and each one of them can be treated in a different way. The problem with such approaches is that they are constrained to domains where the structure of documents is common and is known in advance.

The above three levels of how a text can be treated are common to IR, Text Categorization and IF. Their instantiation however, depends on the particularities of each application and the goals it is trying to achieve. The choice also depends on the way the information class is represented and the corresponding similarity measure between the two. We will also see that the choice of a semantically richer representation does not imply improved performance. On the contrary, the more semantically rich the used representation is the less flexible the overall system becomes.

4.1.2. Representation of the Information Class.

The representation of the information class that the evaluated documents are matched against, depends on the technology and its fundamental goals. In IR for example, the user's information need or interest is usually expressed as a query formed by the conjunction and disjunction of user specified terms (Boolean query). This choice of representation is implied by the fact that IR is mainly concerned with satisfying a one-time user information need. Thus the user has to be able to express his need in a straightforward way and to have full control over the representation. The disadvantage of this simple and comprehensive way of expressing an information need is that queries can not perform any actions on their own. Their effectiveness depends on the analysis of the collection of documents that usually produces an *inverted index*, i.e. a data structure that lists for each word in the collection all documents that contain it and the frequency of occurrence in each document. Furthermore, despite the use of query expansion algorithms that appropriately modify a user's query according to his feedback to the initial results, a query is just a temporary representation that is discarded after the end of the information seeking episode.

Text Categorization and IF use more elaborate representations. Both technologies use some representation of the information class that in the first case corresponds to one or more topic categories while in the second case it corresponds to the interests and needs of the user. As already mentioned, the basic difference between the two is that Text Categorization is concerned with relative static categories, as in the case of categorizing magazine articles according to some pre-specified categories (Moens, M. and Dumortier, J. 2000). In most such cases a large enough training set of pre-categorized stories is available and is used for training a classifier using *machine learning* techniques (Chen, H. 1995, Cohen, W. W. and Singer, Y. 1996, Lewis, D. D. and Ringuette, M. 1994). The machine learning algorithm used usually also implies the produced representation. *Decision Trees* for example are induced using machine learning algorithms like *ID3* and *C4.5* (Mitchell, T. M. 1997) while *back propagation* and *Hebian learning* are used to train *Neural Networks* (Chen, H. 1995, Haykin, S. 1999). Irrespectively however of the machine learning algorithm used, the produced classifier is relatively static like the category or categories that it represents. After a classifier is successfully trained then it is used in the real situation for document classification. If the application domain changes this usually implies re-training the classifier or training a new one. In IF however, a user profile can be seen as a classifier that has to be able to constantly adjust to the changing topic categories of interest to the user. This need has resulted in the emergence of the research fields of adaptive information filtering and personal assistants (intelligent agents). We will discuss these two areas of research in IF in a following section.

4.1.3. Similarity Measure.

The similarity measure used to assess the relevance of a document to the information class depends on both the representation of the document and the representation of the information class. We can however distinguish between two basic models. According to the *exact match* model a system returns all the documents that precisely satisfy some structured information need expression (e.g. query or rule). This approach is used basically by IR systems and its disadvantage is that usually the results set is either empty or huge and unwieldy (Manning, C. D. and Schutze, H. 1999). The second model ranks documents according to their *estimated* relevance to the information class. The similarity measure assigns a value to each one of the evaluated documents. This value corresponds to the documents estimated relevance and can be used to rank documents by decreasing order of relevance. In the case of text categorization a threshold is used to decide if a document is going to be assigned to a certain category or not. In the same way a threshold can be used by IF systems to decide if a document is going to be presented to the user or not. The use of a threshold is not however mandatory. *Routing* systems for example, a special case of IF systems, present all the incoming documents to the user by decreasing order of relevance without removing any documents from the incoming stream.

4.2. Term Weighting.

4.2.1. Dimensionality reduction.

Term weighting is one of the most important components of all these information access technologies. Studies on IR has revealed that the use of weighted, as opposed to, binary content identifiers for document indexing improves the retrieval operations (Salton, G. 1973). The weight of a term also defines its relative importance as part of a classifier or as part of a user profile. Finally, term weighting is also important for automatic dimensionality reduction. Even a moderate-sized text collection can include tens or hundreds of thousands of unique words. The high dimensionality of the feature space is a problem for most *machine learning* algorithms (Yang, Y. and Pedersen, J. O. 1997). One way to reduce dimensionality is to remove non-informative terms. This can be achieved for example by removing words that appear in a *stop list* of "grammatical" or *function words* like, *the*, *from*, and *could*. Although these words are semantically important, and especially in combination with other words, their discrimination power is limited. According to Zipf's law a stop list that covers a few dozen words can reduce the dimensionality of the space by half (Manning, C. D. and Schutze, H. 1999). Term weighting methods however, go one step further. They assess the informativeness of a word and thus enable the selection of the most important terms (*concept terms)* and the removal of the rest from the feature space. The next paragraph presents some term weighting methods and discusses their potential use for IF applications. Another way dimensionality reduction can be achieved is by combining less informative terms to construct new more informative terms in higher-level orthogonal dimensions. *Latent Semantic Indexing* (LSI) achieves exactly this by projecting co-occurring terms into the same dimensions of a "latent" semantic space. For more details on LSI see (Faloutsos, C.

and Oard, D. W. 1996, Manning, C. D. and Schutze, H. 1999). Finally, dimensionality reduction is also achieved using *stemming*. Stemming algorithms like those developed by Porter and Lovins truncate a word into its stem. Words for example like *laughing*, *laugh*, and *laugh* are all truncated into their common stem *laugh*-. Church has studied the use of stemming and although he agrees with previous experimental results showing that stemming does not affect retrieval performance, he concludes that the use of stemming is justified especially for concept terms (Church, K. W. 1995). This suggestion makes stemming a good complementary method for dimensionality reduction.

4.2.2. Term Weighting Methods.

The basic information used for the weighting of a term is:

a) *Term frequency* (*tf*), i.e. the number of times the term appears in a certain document. Term frequency can be used as a measure of the importance of a term within a certain document. The underlying idea is that a term related to the document's topic will appear more frequently in the document than most non-related terms. Functions like the log_2 of the frequency or its normalization to the document's length are also frequently used, although Singhal argues that document normalization reduces retrieval performance (Singhal, A., et al. 1996).

b) *Document frequency* (*df*), i.e. the number of documents in the collection that contain the term. Document frequency reflects the importance of a term within the entire collection. It is based on the assumption that concept terms will only be concentrated to some of the documents in the collection and not evenly distributed among all documents.

c) *Collection frequency* (*cf*), i.e. the number of times the term appears in the complete collection. This is another measure that reflects the importance of a term within the whole collection. The assumption here is that very frequent terms are more likely to be function terms and not concept terms. At the same time the discrimination power of very rare words is also limited. Many more elaborate functions have been developed for measuring the importance of a term within a collection. The next paragraph presents some of them.

We can distinguish between two types of term weighting methods. *Task-specific* methods take into account information about pre-assigned document categories. In contrast *task-independent* methods do not take into account any such information and just base the evaluation of a term on its general statistics in the collection. More specifically:

Task-specific term weighting methods use relevance information to distinguish between documents that are relevant to a topic category and those that are not. The statistics of a word can then be represented by a two way contingency table of a term *t* and a category *c* (table 1). In this table, *R* is the number of relevant documents, *I* is the number of not relevant documents, *r* is the number of relevant documents that include the term, *s* is the number of not relevant documents that include the term and finally *N* is the total number of documents in the collection.

	Relevant Documents	Irrelevant Documents	
Includes term	*r*	*s*	*F=r+s*
Term is not included	*R-r*	*I-s*	*N-F*
	R	*I*	*N*

Table 1. Term statistics using relevance information.

One of the most used task-specific term weighting method is the χ^2 (*chi-square*) test. χ^2 was initially introduced for assessing the dependence between two terms (Manning and Schutze, 1999). We can thus identify collocations between terms. Analogously, it can also be used to assess the dependence between a term and a topic category. Based on the values in table 1 we can calculate the difference between the observed frequencies and the frequencies expected according to the hypothesis of independence (formula 1). If the difference is large enough then we can reject the null hypothesis of independence. The χ^2 test is used and/or tested by (Moens, M. and Dumortier, J. 2000, Ng, H. T., et al. 1997, Yang, Y. and Pedersen, J. O. 1997).

$$\chi^2 = \frac{N \times \left(r\left(I-s\right) - \left(R-r\right)s\right)^2}{\left(r+\left(R-r\right)\right) \times \left(s+\left(I-s\right)\right) \times \left(r+s\right) \times \left(\left(R-r\right)+\left(I-s\right)\right)} \qquad \textbf{(Formula 1).}$$

A variant of the χ^2 is proposed by (Ng, H. T., et al. 1997). The metric proposed by Ng is called *correlation coefficient C*, where $C^2=\chi^2$. The difference between the two metrics is that, while the χ^2 test also identifies terms that are indicative of a document's non-membership to a topic category *c*, the correlation coefficient metric only picks terms that are indicative of membership. In other words, only terms from relevant documents are selected. According to Ng, the experimental results indicate that the use of the latest kind of terms (local dictionary) results in improved categorization performance.

An extensive comparison of term weighting methods is presented by (Yang, Y. and Pedersen, J. O. 1997). Among the five weighting methods compared, three of them are task-specific: *Information Gain*, *Mutual Information* and χ^2. Information gain is a measure motivated by information theory. It measures the number of bits of information gained for the category prediction if we know the presence or absence of a term in a document. Mutual information (MI) is a similar information theoretic metric that measures the reduction in uncertainty of one random variable due to knowing about the other (Greiff, W. R. 1998, Manning, C. D. and Schutze, H. 1999). Based on the values in table 1, MI is defined as (Formula 2).

$$MI = \frac{r \times N}{\left(r+\left(R-r\right)\right) \times \left(r+s\right)} \qquad \textbf{(Formula 2.)}$$

Finally, *Term Precision* and *Relevancy Score* are two more task-specific term weighting methods. Term precision is analyzed by (Yu, C. T., et al. 1982), where it is also related to the term's frequency of occurrence. The relevancy score metric is introduced by (Wiener, E., et al. 1995). It is used to measure how "unbalanced" a term is across documents related to or not related to a specific topic category *c*.

Task-independent term weighting methods do not take into consideration any relevance information and just base the assessment of a term's goodness on its frequency characteristics. Essentially, this type of methods implicitly or explicitly measure the deviation of a term's distribution in a collection from the expected distribution of a random, non-content term, as it is usually expressed using the Poisson distribution (Manning, C. D. and Schutze, H. 1999, Robertson, S. E. and Walker, S. 1994).

The simplest way to weight a term in a collection without using relevance information is to measure its document frequency (*df*) or even better its *inverse document frequency* (*idf*) which in its simplest form is defined as *N/df*. Despite its simplicity *idf* gives good results and has also the advantage that it easily scales to very large corpora (Harman, D. 1986, Moens, M. and Dumortier, J. 2000, Yang, Y. and Pedersen, J. O. 1997). It is also interesting to note that according to both Yang and Greiff, there is a correlation between *df* (or *idf*) and the more informational rich measure of mutual information (Greiff, W. R. 1998, Yang, Y. and Pedersen, J. O. 1997). An alternative to *idf*, called *Residual idf*, was introduced by (Church, K. W. 1995). Residual *idf* can be defined as the difference between the logs of a term's actual document frequency and the document frequency predicted by the Poisson distribution.

Term Discrimination Value is another task-independent term weighting metric described by (Yu, C. T., et al. 1982). It is based on the hypothesis that a good content term is one, which decreases the density of the document space. Reducing the density of the document space makes it easier to distinguish a document in a collection from its neighbors. We can thus calculate a term's discrimination value as the difference S_b-S_a, where S_b (S_a) is the density of the space before (after) assigning the term to the documents in the collection.

Two more task-independent term weighting methods can be used to assess the importance of a term within a collection. The *Term Strength* metric is described and evaluated by (Yang, Y. and Pedersen, J. O. 1997). Given a pair of related documents, a term's strength can be calculated based on the estimated conditional probability that a term appears in one of the documents given that it appears in the other. Pairs of related documents can be constructed out of the collection using the cosine similarity measures. The second measure, *noise*, is referenced by (Harman, D. 1986). It measures the concentration of a term and is calculated using the following formula (Formula 3.), where tf_{dk} is the frequency of term *k* in document *d* and cf_k is the term's collection frequency.

$$noise_k = \sum_{d=1}^{N} \frac{tf_{dk}}{cf_k} \log_2 \frac{cf_k}{tf_{dk}} \qquad \textbf{(Formula 3.)}$$

The term weighting methods described so far assess the importance of a term within the complete collection. The simplest way to estimate the importance of term within a document is by using its *tf* or the $\log_2$ of *tf* in the document (Harman, D. 1986). Alternatively we can also normalize *tf* to the number of words in the document or to the maximum number of times a term appears in the document (Moens, M. and Dumortier, J. 2000). The most interesting finding however is that metrics that assess the importance of a term within a collection and metrics that assess the importance of a term within a document can be combined either additively or multiplicatively. In both cases performance is improved over the use of a single method (Harman, D. 1986). We should however note that any such combination of metrics is actually a new metric of the importance of a term within a document.

The most frequently used combination of metrics is the *Term Frequency Inverse Document Frequency* (*tfidf*) weighting method. It combines *tf* and *idf* to measure the within document importance of a term (*tf*) by also taking into account the term's importance within the complete collection (*df*). The *tfidf* metric has been extensively used by many information access systems, usually combined with a cosine similarity measure. Different variants of *tfidf*, which incorporate normalization and/or logarithmic dumping of the effect of the two parameters (*tf* and *df*) can be used (Salton, G. and Buckley, C. 1988). The disadvantage however of any within document importance metric is that a term's weight is related to a specific document and not some general topic category that includes many documents.

4.2.3. Term Weighting and Personalized Information Filtering.

All of the above term weighting methods were introduced as part of research in IR and Text Categorization. As we will see in the next section, IF applications adopt these methods to create weighted representations of the content of documents or to select the most important terms to be added in a user's profile. To enable the use of task-specific term weighting methods most IF systems expect that the user will give feedback to the system's suggestions by identifying both relevant and non-relevant documents to his interests. We however argue that it can not be guaranteed that a user will give feedback on non-relevant documents. In a hypothetical scenario and given that an IF system performs satisfactorily, it is likely that a user will find that the system's suggestions satisfy his information need and will thus give only positive feedback. For example a system presents to the user a list of documents ranked by decreasing estimated relevance. The user reads the first three documents and having found what he needs he does not go on to read the rest of the presented documents. As a result the user will only give positive feedback for the first three documents and no negative feedback at all.

We thus believe that a successful system that provides personalized information filtering should be able to weight terms using only positive feedback. This requirement excludes the use of most of the task-specific methods presented above, due to their dependence on information on both relevant and non-relevant documents. The use of task-independent methods is not the obvious solution. These methods do not use as much information as their task-specific counterparts and so they are usually less efficient. It is one of our research goals to investigate the development of term weighting methods that use only positive feedback and are in general specialized to the task of IF and not just adopted from IR or Text Categorization.

5. APPROACHES TO INFORMATION FILTERING

As we have discussed in the previous section IF usually involves the analysis of a document and as a result its representation based on its actual content. The exception is the use of conceptual hierarchies like thesauri (Bloedorn, E., et al. 1995) and WordNet (Mock, K. J. and Vemuri, V. R. 1997). Even in these cases however the indexing of documents based on these abstractions can be done automatically. We can thus conclude that the first of the requirements of IF systems for KM that we have mentioned in the previous section is satisfied by most IF systems. In the rest of this section we will

concentrate on the last two requirements and we will see that existing IF system do not satisfy both satisfactorily enough.

5.1. Learning the user information interests or needs.

An IF system should be able to learn from the user and thus minimize his interaction with the system. This usually implies the use of machine learning techniques for constructing a user profile based on documents that the user specifies as relevant or non-relevant. In most of the cases the machine learning algorithms used are adopted from Text Categorization. We discuss the use of machine learning algorithms by IF systems without getting into the details of the algorithms or any extensive comparison between them.

The *InfoFinder* IF agent for instance, uses the *ID3* machine learning algorithm to induce a decision tree based on HTML documents that the user specifies as relevant or non-relevant while browsing the Internet (Krulwich, B. and Burkey, C. 1997). The selected documents are first categorized by the user and then analyzed using syntactic heuristics, like capitalization or Italics, to extract significant phrases that represent the document's content in a semantically rich way (see previous section) (Krulwitch, B. 1995). The extracted features that represent each one of the documents are then used by the machine learning algorithm to construct a decision tree for each one of the user-specified categories. Queries are formulated based on this tree and are submitted to known search engines. The documents returned by the search engines are filtered using the decision tree and those that the user gives feedback on are collected and used for revising the tree. Tree revising does not involve restructuring but only its extension with more specialized leafs. One problem with this approach is that it is based on the assumption that authors will be using the syntactical patterns on which feature selection is based.

A semantically different representation of documents is used by (Bloedorn, E., et al. 1995). Documents are represented using summery level features based on a thesaurus and with the help of the Subject Field Coder (SFC). This representation of documents is combined with features related to People, Organizations, and Locations (POLS) and keywords weighted using *tfidf*. Once more the user provides the system with positive and negative document examples which are used to generate rules for future filtering of information. Two rule generating machine learning algorithms have been evaluated. *AQ15c* generates disjunctive normal form (DNF) expressions and *C4.5-Rules* generates rules based on the decision tree induced by the *C4.5* algorithm. One problem with this approach can be the use of rules for representing the user interests. Although rules are generally easily comprehensive by users, they are nevertheless not as credible as other representation like linear classifiers or prototypes generated using genetic algorithms, because users can find counterexamples to any rule (Pazzani, M. J. 2000). More details on the use of genetic algorithms for the construction of prototypes for information filtering can be found in (Chen, H. 1995).

Another rule induction algorithm is *CN2* used by the *Magi* (Mail Agent Interface) mail filtering system (Payne, T. R. and Edwards, P. 1997). The problem with *CN2* is that it expects single values for each one of the features in a training instance. More than one instances have thus to be formed by each mail message used for training the system. This

results in an increased number of training instances and hence to an increased training time. Alternatively, Payne proposed the use of an instance-based learning algorithm called *IBPL1* and its improved variant *IBPL2*. Instance-based learning algorithms learn by storing complete instances as points in the feature space and classify new instances by comparing them to the memorized instances. In this sense *IBPL1* can be seen as an extension to *Memory Based Reasoning* (MBR). It uses the *k-nearest* algorithm to classify a new instance according to the *k* closest memorized instances. The advantage of *IBPL1* and *IBPL2* over *CN2* is that multiple values are allowed for the different features. The distance between two instances can then be calculated either by summing up the calculated distance between all possible combination of features and then averaging the sum by the number of combinations (*IBPL1*) or using the *closest value* distance that was found for the possible combinations (*IBPL2*).

MBR is also used by the *ABIS* (Agent Based Information System) (Amati, G., et al. 1997). *ABIS* support both the filtering of documents returned by a user query (*query mode*) and the active search of an archive (*surfing mode*), where the agent autonomously navigates the network in a quest for relevant documents. In *ABIS* a user profile consists of three components. The *Constraint Set* maintains descriptions of general user preferences like the number of documents to be retrieved. The *Personal List* contains a number of URLs used as starting points for the system's surfing mode. Finally the *Preference Profile* is the actual description of the user's interests. The representation of documents, queries and the user's profile is handled by the *Harvest* public-domain program and a refined form of the *SOIF* format (Summary Object Interchange Format) that *Harvest* is using. According to *SOIF* each document is represented by a number of attributes, each one containing a summary of the document's actual content that corresponds to the attribute (e.g. title, author, abstract, etc.). Each document can thus be represented by a list $d=(< a_{d1}, u_{d1} >, \ldots , < a_{dn}, u_{dn} >)$. Queries can also be formulated as conjunctions and disjunctions of such lists. Finally, the user profile consists of three different structures: the *Preference Profile*, the *Situation Set* and the *Constraint Set* that we have already mentioned. The *Preference Profile* is represented in the same way as a query. The *Situation Set* is a database of previous interactions with the system on which MBR is based. Each previous interaction is represented by a 4-tuple of the form $S=<q, d_i, act_i, E_i>$, where q is a query or the *Preference Profile*, d_i is a retrieved document, act_i is the action performed on the document by the user and finally E_i represents data for lower level functionalities. Based on the memorized interactions the system can infer an action to be performed on a retrieved document. Retrieved documents are either rejected or accepted by the user and according to the action that he performs on them new situations are created and added to the database. In this way the profile is updated. One problem with *ABIS* is that to initialize his profile a user has to fill several forms by specifying keywords for different attributes. In this way *ABIS* inherits some of the disadvantages of knowledge-based approaches to KM that we have already discussed.

In the same paper (Amati, G., et al. 1997), Amati also presents another IF system, *ProFile*. *ProFile* specializes in the filtering of netnews, but it can also be extended to the filtering of other incoming information streams. To initialize his profile a user first specifies a number of conceptual classes and assigns a number of keywords to each one of them. This initial description of the user interests is used by the *FIFT* service (Fub Information Filtering Tool), which is a customized version of the *SIFT* filtering system

developed by Stanford university, to filter out a set of documents from a collection. The user then evaluates some of the documents in this set using a scale from 0 to 10. Learning takes place using both relevant and non-relevant documents on the basis of the *RSJ* probabilistic model introduced by Robertson and Spark Jones in 1976. The model was extended to use weighted and not only binary features. Feature weights are calculated using the *tfidf* method and the result of the learning process is a weighted feature vector representing each one of the categories of interest to the user. The similarity between the profile and a document is calculated using the inner product of the corresponding weighted vector representation. Whenever the user evaluates a new document the profile is updated by reconstructing the corresponding weighted vector.

A similar approach is followed by the WWW browsing assistant *Syskill & Webert* (Pazzani, M. and Billsus, D. 1997). In this case for each one of the user specified topic categories, the user supplies the URL of an index of pages relevant to the topic. Binary features are extracted from the indexed pages using the information gain metric and a naïve Bayes classifier is used to construct the user profile. The Bayes classifier is also used to update the profile according to the user's positive and negative feedback. The algorithm used was compared to the nearest neighbor, PEBLS, ID3, Rocchio's and the backpropagation (Neural Networks) learning algorithms, and none of them performed as well. Based on the results Amati argues that in filtering documents there is not a need for weighted feature instead of binary and for non-linear classifiers instead of linear do. The results however also reveal that IG feature selection method does not perform as well as expected. Alternatively the use of user selected terms and the incorporation of WordNet is also investigated.

As already mentioned WordNet is also used by (Mock, K. J. and Vemuri, V. R. 1997). As part of the *INFOS* filtering system (Intelligent News Filtering Organizational System), a WordNet based *Case-Based Reasoning* (CBR) component is used to complement the *Global Hill Climbing* (GHC) method whenever the latest can not classify a retrieved document. Previously classified documents are indexed using WordNet concepts so that they can be matched to a retrieved document. A retrieved document is then classified to the category of the closest memorized document. The complementary use of Hill Climbing and CBR resulted in a higher correct classification percentage than GHC. However the percentage of correct classification was still quite low (60%). According to Mock, this low performance is due to the influx of new topics that *INFOS* has not yet learned to classify, and to changes in the user interests.

Finally, we review the approach followed by (Soltysiak, S. J. and Crabtree, I. B. 1998). A filtering system that tries to minimize the user's involvement in the construction of his profile is proposed. The system uses *Prosum*, a text summarizer, to present each document as a vector of weighted terms and phrases. Initially the user supplies the system with a number of WWW pages and e-mails. These documents are analyzed by *Prosum* and then they are clustered based on their weighted vector representation to produce interest clusters. These interest clusters constitute the initial user profile. Retrieved documents are also first clustered and then compared to the existing clusters in the profile to assess their similarity. If a cluster of retrieved documents is close enough to an existing cluster then the two clusters are merged to form a new cluster. In the opposite case that a cluster of retrieved documents is not close enough to any stored clusters then the user is asked if he wants to add the cluster in the profile as a negative or positive example. The

characteristics of the different clusters like the size, the age or the mean document size are also used to form heuristics for automatically classifying clusters as positive or negative.

5.1.1. Learning vs. Adjusting.

The IF systems presented above are only some of the existing IF systems that use machine learning algorithms for learning a user profile based on positive and negative document examples that the user supplies. Theoretically, any machine learning algorithm used for text categorization can be appropriately adapted to support learning of user profiles. Another example is the EG algorithm (Callan, J. 1998). We however argue that in addition to any individual disadvantages of the above approaches, the use of such machine learning algorithms has inherent drawbacks.

First of all, these algorithms are designed for increased accuracy when trained with a sufficiently large number of both positive and negative examples. The amount of documents that a user can initially provide is usually not sufficient and in addition, as already mentioned it is not guaranteed that the user will provide negative feedback. The most significant disadvantage however is that the user profile that these algorithms generate cannot adapt to changes in the user interests as identified by Mock (see above). Given the feature or document space (in the case of Instance-Based Learning or MBR), the above machine learning algorithms initially learn a definition of the user profile as an area in this space (profile initialization). When after initialization the user supplies the system with more positive or negative documents this area just becomes more *concrete* or *larger* (figure 4.). In the case for example of decision trees more leafs are added to the existing tree. If MBR is used then more examples are added to the existing memorized examples. This disadvantage is also revealed by the fact that in most of the cases the user has first to define different topic categories and a distinct profile is built for each one of them. The design choice of such machine learning algorithms could thus be justified for a strict IF application where the assumption of long term user information interests maybe holds true. This is however not the case in the domain of knowledge management where the user's information needs change quite quickly based on his task at hand and what remains relatively stable is maybe his expertise or his general social context. As already mentioned in the section on KM, an information filtering system can be suitable for the development of a KM application only when it is able to adjust to the changes that the circumstances imply to the user's information needs.

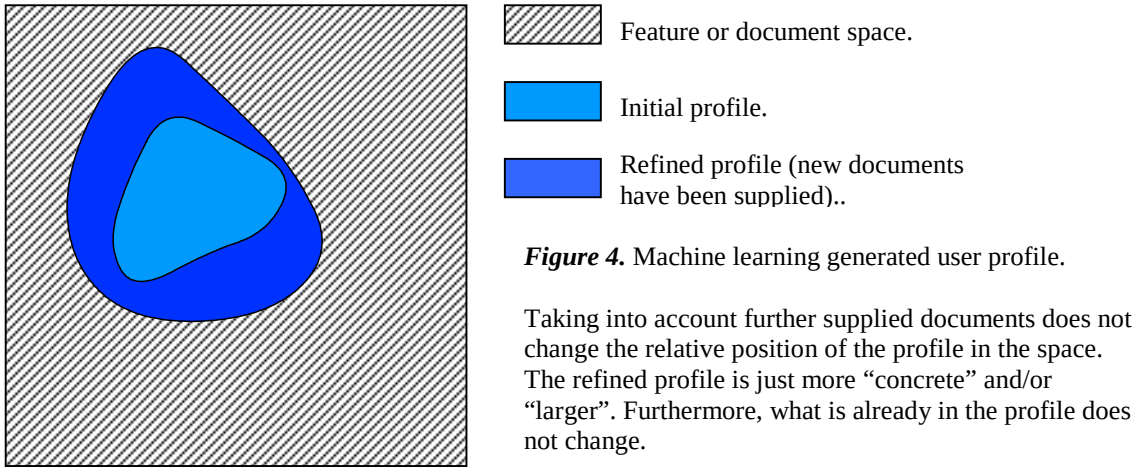

Figure 4. Machine learning generated user profile.

Taking into account further supplied documents does not change the relative position of the profile in the space. The refined profile is just more "concrete" and/or "larger". Furthermore, what is already in the profile does not change.

5.1.2. Adjusting requires both adapting and evolving.

So far in this document we avoided the use of the word “adapt” to describe the required ability of a KM oriented IF system to adjust to the changes in the user’s information needs. This is because we believe that this requirement implies an IF system that is capable to both *adapt* and *evolve*. It is thus pertinent at this point to clarify the way these two concepts are used in the rest of this report.

We characterize a representation of the user information needs *adaptive*, when it is able to reform itself in accordance to the user feedback. This is not only achieved by the addition of new features or instances to the existing representation structure but also by the appropriate rearrangement of the interrelations between the structure’s components and/or of their relative importance as part of the structure. These additional degrees of freedom allow for the definition of the user’s information needs as an area in the feature space that has the ability to reshape itself although the underlying structure’s components can remain the same (figure 5.). Changes to the user’s information needs can be quickly reflected by this representation structure even based on the user’s feedback on just one document. However, as the relative approaches will reveal, only modest changes to the user’s information needs can be reflected by this kind of representation. The reshaping of the user’s information needs area takes place locally (figure 5.). The representation cannot follow a complete change of the user’s information needs in the feature space.

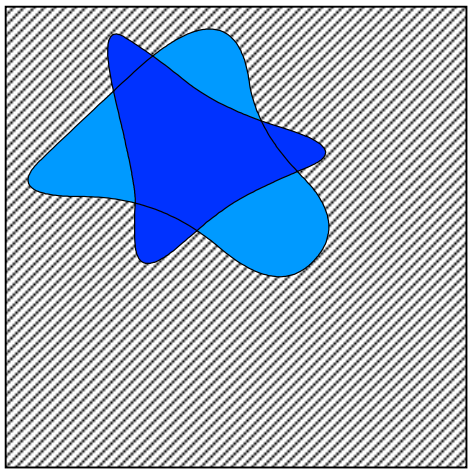

Figure 5. Adaptation of a user profile.

Taking into account further supplied documents can reshape the area of interest to the user in the feature space although the components of the underlying representation structure can remain the same. This reshaping takes place locally.

We can now characterize a representation of the user’s information needs *evolving*, when it is able to follow changes to these needs anywhere in the feature space and is not anchored to some certain area by its initialization. As it will be presented later on, evolving information filters maintain for each individual user, a population of profiles that is spread around the feature space. Those profiles that cover areas of interest to the user survive while those that don’t are purged. Genetic operations are used to generate new profiles that either search the neighboring area to the fittest profiles (crossover) or arbitrary explore new areas in the space (mutation). When the user’s information needs change radically or a new area of interest arise, one of the generated profiles eventually covers this new area and its fitness increases (figure 6.). This reaction is not immediate

and usually a number of system interactions with the user are required for a profile to converge to the new area of interest.

Conclusive, there is a trade off between the rapid adjustment achieved at a local area through adaptation and the progressive adjustment to radical changes in the user's information needs that evolution accomplishes. Ideally, a KM oriented IF system should be able to both adapt and evolve. Adaptation would allow the system to adjust to short term changes in the information needs, like those that occur in the context of a knowledge intensive task, when for example a document is received or when the user has communicated with a peer. On the other hand radical changes can occur when for example the user starts working in a new project. An evolving user profile could then adjust to any such changes. The following sections present adaptive IF systems, evolving IF systems and systems that to some extent combine both adaptation and evolution.

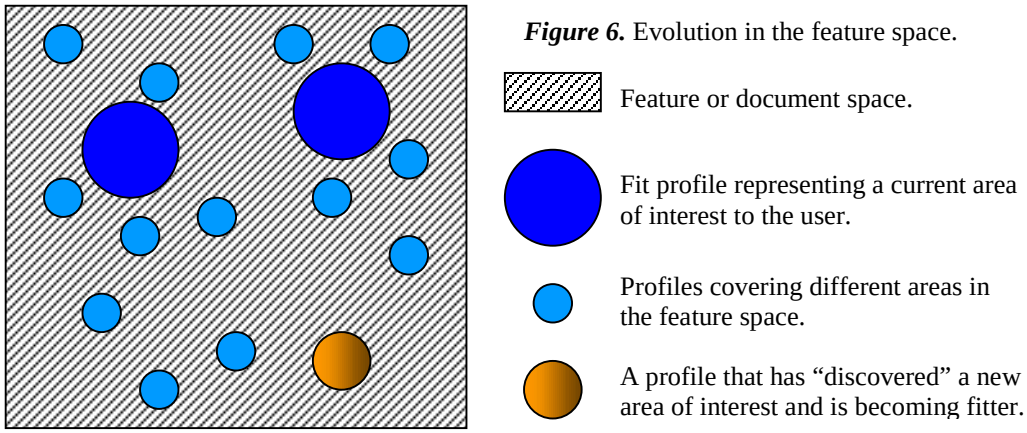

Figure 6. Evolution in the feature space.

5.2. Adaptive Information Filtering Systems.

To demonstrate adaptation in IF we present two IF systems, *INFOrmer* and *PSUN*. Both systems use a *connectionist* representation of the user's interests, which is constructed as a network of interconnected features. Connectionism has also been investigated in IR where for example a network of documents, keywords and author names can be appropriately formulated to support the adaptive retrieval of documents (Belew, R. K. 1989). The difference here is that the constructed network is common for all users while in IF a personalized representation of the user's interests is built for each individual user.

More specifically, in *INFOrmer* an associative network is used to represent the user's profile (Sorensen, H., et al. 1997). To construct this network the user specified documents are first preprocessed. This phase involves the removal of stop words, stemming of words with common root and the identification of sentences in the text. No term weighting mechanism is used to identify concept terms. An analysis of the preprocessed documents follows. Each term in a document is represented in the profile by a weighted node. Adjacent terms are linked together with weighted edges taking into account the boundaries of sentences. Only terms that appear in the same sentence are linked together. Initially all nodes and edges have the same weight. This first representation of the user's profile is then modified by merging together nodes representing the same term. As a result the weight of nodes and edges is also

appropriately modified. This has the effect that nodes and corresponding phrases that appear frequently in relevant documents have increased weight. A similar representation is also used for incoming documents with the difference that nodes and edges have no weights. The similarity between the profile and the incoming documents is based on the structural similarity between the corresponding graphs. To compare the profile to a document a spreading activation function is used. An activity is placed in those nodes (terms) in the document that also appear in the profile and is then appropriately leaked to neighboring nodes. The result is that nodes comprising phrases in the document that also appear in the profile are highly activated. The number of nodes with activation that exceeds a certain threshold is used as a relevance metric. For more details on this process see (O'Riordan, A. and Sorensen, H. 1995). Adaptation of the profile is based on the user feedback. Highly activated terms in a relevant document are either used to reinforce the same terms in the profile or are added to the profile if it doesn't include them.

A similar approach is followed by the *PSUN* (Profiling System for USENET News) IF agent (Sorensen, H. and Mc Elligott, M.). *PSUN* is also based on the formation of phrases as orderly linked terms. It's theoretical motivation is Schank's "scripts" and Minsky's "K-lines" (Mc Elligott, M. and Sorensen, H. 1993). In *PSUN* a two layer representation of the user's interests is used. At the first layer and as in the case of *INFOrmer*, nodes (terms) are appropriately linked together to form phrases. Stop word removal and stemming however, is not applied. Another difference is that a dual weighting mechanism, based on a combination of constrained and unconstrained Hebbian learning, is used. Words are weighted in accordance to their frequency in the user-specified documents. Edges however are weighted in two ways. A *weight* is assigned to each edge to account for its local significance. At the same time to measure the global contribution of the phrase within the profile a *strength* is also assigned to each edge. This is achieved by initially providing each node in the profile with a fixed amount of strength points. Unconstrained Hebbian learning is then used to share these points among the parent node's outgoing links. When the strength that the parent node can provide is depleted, links have to compete with each other for their strength points. Constrained Hebbian learning is used in this case. The overall effect of this process is that links that represent insignificant phrases eventually get "weaker" and thus "forgotten". For details on the learning mechanisms used see (Mc Elligott, M. and Sorensen, H. 1994). Now at the second layer a *supervisor* is formed for each pair of terms that are strongly related. Each supervisor is allotted a contribution according to the strength of its corresponding term pair. The contributions of all the supervisors in the profile add up to a specific number (100). When evaluating a document a supervisor is responsible for looking in the text to find the pair of words that it represents. If the pair is found, the supervisor fires its contribution. The document's relevance can be now calculated as the sum of the contributions of the supervisors that have fired. Once more the adaptation of the profile is based on the user feedback. According to the documents that the user has deemed relevant the weights and strengths of the nodes and edges in the profile are appropriately modified and the supervisors are reformed.

The similarity between the two approaches is obvious. One of their strengths is the ability to represent the dependency between terms without at the same time being constrained by having to store complete phrases. The linking between terms can represent a number of alternative phrase expressions. As a result the constructed profile constitutes

a powerful nonlinear classifier. At the same time the representation is flexible. The importance of the profile's components (nodes) and their interrelations (edges) can be rearranged to appropriately reshape the area of interest to the user in the feature space (adaptation). However, in both cases there is not a mechanism for removing terms in the profile that no longer represent concepts of interest to the user. Although in the case of *PSUN* the problem is alleviated to some extent by the ability of the profile to "forget" phrases (links between terms), terms remain in the profile and in addition to occupying storage space they anchor the representation to some area in the feature space. This is evident by the fact that in both cases the authors suggest the use of different profiles for different, not overlapping user interests.

5.3. An Evolving Information Filtering System.

As already mentioned the ability of a representation of the user's interests to move around the feature space can be achieved with evolving mechanisms. Evolutionary IF systems maintain a population of profiles for each individual user. The genetic operations of crossover and/or mutation are used to generate new profiles that search the feature space in an arbitrary but at the same time directed way.

In *Amalthaea* an artificial multi-agent ecosystem of evolving agents that cooperate and compete is used (Moukas, A. and Maes, P. 1998, Moukas, A., et al. 1999). The ecosystem is composed by two general species of agents, namely Information Filtering Agents (IFAs) and Information Discovery Agents (IDAs). Competition for survival takes place among agents of the same species, while co-operation is performed between agents of different species, explicitly for their own sake and implicitly to enhance the systems performance. Each IFA has a chromosome composed mainly by a weighted keyword vector. IFAs thus specialise to a certain domain of the user's interests. Collectively they form a representation of the user's interests with such a diversity that makes it flexible enough to adjust to any changes. IFAs use their keyword vectors to filter the documents returned by the IDAs. IDAs are responsible for information resource handling. They act parasitically on existing search engines to find and fetch the actual information that the user is interested in. In the co-operation between an IDA and an IFA, the IDA uses the IFA's keyword vector to form a query in the most appropriate way for the search engine it specializes in. The returned documents are passed to the IFA, which filters them and returns the most appropriate one to the user. *Amalthaea* thus combines the information filtering performed by the IFAs with the parasitic information retrieval functionality of the IDAs. The evaluation of the agents takes place on the base of an economic model. IFAs receive the user's positive feedback in terms of credit and then give some of this credit to the IDAs they cooperated with to find the relevant documents. All agents pay "rent" to inhabit the system's environment and as a result those that do not perform well run out of credit and are purged from the environment. The two species are evolved separately. Only the fittest agents of the two populations are "mated" using crossover, to produce new hopefully better offspring. Mutation is also used to explore new areas in the feature space.

The basic problem with *Amalthaea* is that no adaptation is performed by the individual profiles during their "life time". Once a profile has been generated the corresponding representation remains as it is and the profile survives as long as this fixed

representation corresponds to a domain of interest to the user. The following systems overcome this problem by providing the individual profiles with the ability to learn locally based on the user's feedback.

5.4. Combining Evolution with Local Learning.

A similar system to *Amalthaea* is *NewT* (Sheth, B. D. 1994). The basic difference between the two systems is that *NewT* uses only filtering agents that filter incoming USENET news articles. No retrieval functionality is thus needed. In addition, although both systems represent individual profiles using weighted keyword vectors with *tfidf*weights, *NewT* distinguishes between different article fields like the author, the title and the article's body. The most interesting however difference is that *NewT's* agents have the ability to learn locally. When an article that an agent has recognized as relevant receives positive feedback by the user, then the agent's keyword vector is moved towards the vector representing the article. The same learning mechanism is also used by (Tauritz, D. R., et al. 1999). Trigrams are however used instead of keywords to form the weighted vectors representing the individual profiles.

Local learning is also exhibited by the evolutionary filtering system presented by (Baclace, P. E. 1991). Baclace avoids the use of weighted keyword vectors and instead uses agents that represent either a single field-value pair or conjuctions of such pairs. Each agent is allotted a bid in the range between [-1,1] similar to the contribution of supervisors in *PSUN*. An agent's bid is appropriately adjusted based on the user's feedback using constrained Hebbian learning. Furthermore, an economic model similar to the one used in *Amalthaea* is used to evaluate individual agents. In the same way agents that run out of money (credit) are either pruned from the profile or maintained in a passive mode as potential parents of new agents. Crossover is used to create new agents by combining the field-value pairs of the parents.

Finally an interesting approach is followed by the *InfoSpiders* IF system (Menczer, F. and Monge, A. E. 1999). In *InfoSpiders* a population of agents is maintained that autonomously search the web in favor of the user. The population is initialized by assigning to each agent a starting web page, an initial amount of energy and a query, which can be the same for all the agents. Each agent browses the network like the user would do, by computing the relevance estimate for each outgoing link from the current document. The agent consumes energy both to visit a new document and send a relevant document to the user. Energy is gained based either on the relevance of the document to the given query or through direct user feedback. Local learning is performed using the connectionist version of Q-learning. A feed-forward neural network is trained based on the difference between the (estimated) relevance of the current document, the estimate of the link that led to it and the corresponding change in energy. Agents are selected for reproduction in a local fashion according to the comparison of an agents energy to a certain constant. The neural networks learned weights are "Lamarckian" in that they are inherited by offspring at reproduction.

The above systems implement a hybridization of genetic algorithms and local learning. Distributed learning of individual agents is combined with the evolution of the population. Individual learning provides a quick acting adaptation mechanism while at the same time evolution allows the search of the feature space on a broad level. In most

of the cases however the representation of the individual profiles supports only linear classification. Evolution plays a more significant role in the ability of the system to adjust. We believe that more powerful adjusting mechanisms can be supported by non-linear classifiers that also have the ability to evolve.

6. CONCLUSIONS

This literature review have investigated the role information filtering technology can play in the development of knowledge management systems. The theoretical foundations of knowledge management revealed that information is an important resource for the creation of new knowledge. We have argued that when the individual working on a knowledge intensive task is provided with relevant enough information then the extra insights that this information provides support him in his decision process and thus his further actions. The result of this "informed" action is the creation of new knowledge.

The importance of information is also reflected on the services provided by user-oriented knowledge management systems. These knowledge management systems use different approaches to assess the relevance of information entities to the user and his context. One basic trend is knowledge-based KM systems. We have however seen that these systems suffer from a number of disadvantages in the way personalization of information delivery is supported. Technological and human-oriented considerations indicate that to support the delivery of relevant enough information a KM system has to be flexible enough to adjust to changes in the individual's information needs as these are implied by his task at hand.

Towards this direction we have investigated the potential application of information filtering technology to support the development of a successful KM system. Some first IF approaches to KM indicate that this is an emerging trend. However, the domain of IF is already vast due to the solution it provides to the significant current problem of information overload. Our investigation of the domain of IF was first based on its relation to the better established domains of information retrieval and text categorization. We have defined the basic concepts and have described the common techniques and methodologies. We then concentrated on the application of machine learning algorithms to the IF task. Most of these algorithms are inherited from text categorization and thus are not appropriate enough for IF, especially when the system is required to be able to adjust to changes in the user's information needs.

Adapting and evolving of user profiles can both individually and in combination provide a system with the ability to adjust to changes in the user's interests or needs. While adaptation supports the quick refinement of a user's profile at a local rate, it can not address radical changes of information interests or needs. Adjusting on a broad level can be achieved with evolutionary mechanisms, which however need a number of iterations before they converge to loci of interest to the user. We have presented both adaptive and evolving IF systems. More interesting however approaches are followed by IF systems that combine adaptation at the individual level, based on local learning, with overall evolution guided by the general system performance. These systems are the best candidates for the introduction of IF in the domain of knowledge management.

7. Bibliography

Abecker, A., Bernardi, A. and Sintek, M. (1999). **Enterprise Information Infrastructures for Active, Context-Sensitive Knowledge Delivery.** *ECIS'99-The 7th European Conference on Information Systems*, Copenhager, Denmark

Abecker, A., Bernardi, A. and Sintek, M. (2000). **Proactive Knowledge Delivery for Enterprise Knowledge Management.** *Learning Software Organizations - Methodology and Applications.*, Guenther Ruhe and F. Bomarius, Springer-Verlag.

Amati, G., D' Aloisi, D., Giannini, V. and Ubaldini, F. (1997). "**A Framework for Filtering News and Managing Distributed Data.**" *Journal of Universal Computer Science.*, 3(8): 1007-1021.

Baclace, P. E. (1991). **Personal Information Intake Filtering**. *Bellcore Information Filtering Workshop*

Belew, R. K. (1989). "**Adaptive Information Retrieval: using a connectionist representation to retrieve and learn about documents.**" *ACM*.

Belkin, N. J. and Croft, W. B. (1992). "**Information Filtering and Information Retrieval: Two sides of the same coin?**" *Communications of the ACM*, 35(12): 29-38.

Bellinger, G., Castro, D. and Mills, A. "**Data, Information, Knowledge, and Wisdom.**" .

Benjamins, R. V. (1998). **Knowledge Management through Ontologies**. *PAKM98-Second Internation Conference on Practical Aspects of Knowledge Management.*, Basel, Switzerland, 5.1-5.12.

Bloedorn, E., Mani, I. and MacMillan, T. R. (1995). "**Machine Learning for User Profiles: Representation Issues.**" .

Borghoff, U. M. and Pareschi, R. (1997). "**Information Technology for Knowledge Management.**" *Journal of Universal Computer Science*, 3(8): 835-842.

Borghoff, U. M. and Pareschi, R. (1998). **Information Technology for Knowledge Management**, Springer Verlag.

Brusilovsky, P. (1996). "**Methods and Techniques of Adaptive Hypermedia.**" *User Modeling and User Adapted Interaction*, 6(2-3): 87-129.

Buckingham Shum, S. (1998). **Negotiating the Construction of Organizational Memories**. *Information Technology for Knowledge Management.*, U. M. Borghoff and R. Pareschi, Springer-Verlag**:** 54-78.

Callan, J. (1998). **Learning while filtering documents.** *21st Annual International ACM SIGIR conference on Research and Development in Information Retrieval.*, 224-231.

Chen, H. (1995). "**Machine Learning for Information Retrieval: Neural Networks, Symbolic Learning and Genetic Algorithms.**" *JASIS*, 46(3): 194-216.

Church, K. W. (1995). **One Term or Two?** *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.*, Seattle, WA USA, 310-318.

Clarke, P. and Cooper, M. (2000). **Knowledge Management and Collaboration.** *Proceedings of the Third International Conference on Practical Applications of Knowledge Management (PAKM2000)*, Basel, Switzerland

Cohen, W. W. and Singer, Y. (1996). **Context Sensitive Learning Methods for Text Categorization.** *Proceedings of the 19th Annual ACM / SIGIR Conference on Research and Development in Information Retrieval.*, 307-315.

Cook, S. D. N. and Seely Brown, J. (1999). "**Bridging Epistemologies: The Generative Dance Between Orgaizational Knowledge and Organizational Knowing.**" *Organizational Science*, 10(4): 381-400.

Davenport, T. H. and Prusak, L. (1998). **Working Knowledge: How organizations manage what they know.** Boston Massachusetts, Harvard Business School Press.

Davies, N. J., Stewart, R. S. and Weeks, R. (1998). "**Knowledge Sharing Agents over the World Wide Web.**" *British Telecom Technology Journal*, 16(3): 104-109.

Domingue, J. and Motta, E. (2000). "**PlanetOnto: From News Publishing to Integrated Knowledge Management Support.**" *Intelligent Systems*, 15(3): 26-33.

Faloutsos, C. and Oard, D. W. (1996). "**A Survey of Information Retrieval and Filtering Methods.**" .

Foltz, P. W. and Dumais, S. T. (1992). "**Personalized Information Delivery: An analysis of Information Filtering Methods.**" *Communications of the ACM*, 35(12): 51-60.

Furnas, G. W., Landauer, T. K., Gomez, L. M. and Dumais, S. T. (1987). "**The vocabulary problem in human-system communication.**" *Communications of the ACM*, 30(11): 964-971.

Glance, N., Arregui, D. and Dardenne, M. (1999). **Knowledge Pump: Supporting the Flow and Use of Knowledge.** *Information Technology for Knowledge Management.*, U. Borghoff and R. Pareschi, Springer Verlag.

Greiff, W. R. (1998). **A Theory of Term Weighting Based on Exploratory Data Analysis**. *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, Melbourne Australia, 11-19.

Grudin, J. (1994). "**Groupware and Social Dynamics: Eight Challenges for Developers.**" *Communications of the ACM*, 37(1): 92-105.

Harman, D. (1986). **An experimental study of factors important in document ranking.** *Proceedings of 1986 ACM conference on Research and development in information retrieval.*, 186-193.

Haykin, S. (1999). **Neural Networks: a Comprhensive Foundation**, Tom Robbins.

Konstan, J. A., Miller, B. N., Maltz, D., Herlocker, J. L., Gordon, L. R. and Riedl, J. (1997). "**GroupLens: Applying Collaborative Filtering to Usenet News.**" *Communications of the ACM.*, 40(3): 77-87.

Krulwich, B. and Burkey, C. (1997). "**The InfoFinder Agent: Learning User Interests through Heuristic Phrase Extraction.**" *IEEE Expert*: 22-27.

Krulwitch, B. (1995). **Learning Document Category Descriptions through the Extraction of Semantically Significant Phrases.** *IJCAI Workshop on Data Engineering for Inductive Learning.*

Lewis, D. D. and Ringuette, M. (1994). **A comparison of two learning algorithms for text categorization.** *Symposium on Document Analysis and Information Retrieval.*

Losee, R. M. J. (1989). "**Minimizing Information Overload: the Ranking of Electronic Messages.**" *Journal on Information Science*, 15(3): 179-189.

Manning, C. D. and Schutze, H. (1999). **Foundations of Statistical Natural Language Processing.**, MIT Press.

Masterton, S. and Watt, S. (2000). "**Oracles, Bards, and Village Gossips, or Social Roles and Meta Knowledge Management.**" *Information Systems Frontiers*, 2(3/4): 299-315.

Mc Elligott, M. and Sorensen, H. (1993). "**An Emergent Approach to Information Filtering.**" *UCC Computer Science Journal*, 1(4).

Mc Elligott, M. and Sorensen, H. (1994). **An Evolutionary Connectionist Approach to Personal Information Filtering.** *Neural Networks Conference '94*, University College Dublin, Ireland

Menczer, F. and Monge, A. E. (1999). **Scalable Web Search by Adaptive Online Agents: An InfoSpiders Case Study.** *Intelligent Information Agents: Agent-Based Information Discovery and Management on the Internet.*, M. Klusch, Springer-Verlag**:** 323-347.

Mitchell, T. M. (1997). **Machine Learning**, McGraw-Hill.

Mizzaro, S. (1998). "**How many Relevances in Information Retrieval?**" *Interacting With Computers*, 10(3): 305-322.

Mock, K. J. and Vemuri, V. R. (1997). "**Information filtering via hill climbing, wordnet, and index patterns.**" *Information Processing and Management.*, 33(5): 633-644.

Moens, M. and Dumortier, J. (2000). "**Text categorization: the assignement of subject descriptors to magazine articles.**" *Information Processing and Management.*, 36(6): 841-861.

Moukas, A. and Maes, P. (1998). "**Amalthaea: An Evolving Multi-Agent Information Filtering and Discovery System for the WWW.**" *Autonomous Agents and Multi-Agent Systems.*, 1(1): 59-88.

Moukas, A., Zacharia, G. and Maes, P. (1999). **Amalthaea and Histos: MultiAgent Systems for WWW Sites and Reputation Recommendations.** *Intelligent Information Agents: Agent-Based Information Discovery and Management on the Internet.*, M. Klusch, Springer-Verlag**:** 293-322.

Newman, B. D. and Conrad, K. W. (2000). **A Framework for Characterizing Knowledge Management Methods, Practices, and Technologies.** *Proceedings of the Third International Conference on Practical Aspects of Knowledge Management (PAKM2000)*, Basel, Switzerland

Ng, H. T., Goh, W. B. and Low, K. L. (1997). **Feature selection, perception learning, and a usability case study for text categorization.** *Proceedings of the 20th Annual international ACM SIGIR Conference on Research and Development in Information Retrieval.*, New York, 67-73.

Nonaka, I. and Takeuchi, H. (1995). **The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation.**, Oxford University Press.

O'Leary, D. E. (1998). "**Knowledge Management Systems: Converting and Connecting.**" *IEEE Intelligent Systems*: 30-33.

O'Leary, D. E. (1998). "**Using AI in Knowledge Management: Knowledge Bases and Ontologies.**" *IEEE Intelligent Systems*: 34-39.

O'Riordan, A. and Sorensen, H. (1995). **An Intelligent Agent for High-Precision Text Filtering.** *Conference on Information and Knowledge Management (CIKM'95).*, Baltimore, MD USA, 167-174.

Ovum (1999). **Knowledge Management Building the Collaborative Enterprise**.

Payne, T. R. and Edwards, P. (1997). **Learning Mechanisms for Information Filtering Agents.** *UK Intelligent Agents Workshop*, Oxford, 163-183.

Pazzani, M. and Billsus, D. (1997). "**Learning and Revising User Profiles: The Identification of Interesting Web Sites.**" *Machine Learning*, 27: 313-331.

Pazzani, M. J. (2000). **Representation of Electronic Mail Filtering Profiles: A User Study.** *International Conference on Intelligent User Interfaces.*, New Orleans, LA USA

Reeves, B. and Shipman, F. (1992). **Supporting Communication between Designers with Artifact-Centered Evolving Information Spaces**. *Computer Supported Cooperative Work*, Toronto, Canada, 394-401.

Robertson, S. E. and Walker, S. (1994). **Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval.** *Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Dublin Ireland, 232-241.

Salton, G. (1973). "**Recent Studies in Automatic Text Analysis and Document Retrieval.**" *Journal of the ACM*, 20(2): 258-278.

Salton, G. and Buckley, C. (1988). **On the Use of Spreading Activation Methods in Automatic Information Retrieval.** *Annual ACM Conference on Research and Development in Information Retrieval.*, Grenoble France, 147-160.

Selvin, A. M. (1999). "**Supprting Collaborative Analysis and Design with Hypertext Functionality.**" *Journal of Digital information* **2000**(24 October).

Shardanand, U. and Maes, P. (1995). **Social information filtering: algorithms for automatic word of mouth**. *Conference on Human Factors in Computing Science.*, 210-217.

Sheth, B. D. (1994). "**A Learning Approach to Personalized Information Filtering.**" *Department of Electrical Engineering*: 54.

Shipman, F. M. and Marshall, C. C. (1999). **Formality Considered Harmful: Issues, Experiences, Emerging Themes, and Directions.** *Computer-Supported Cooperative Work.*, 333-352.

Shipman, F. M. and McCall, R. (1994). **Supporting Knowledge-Base Evolution with Incremental Formalization.** *Human Factors in Computing Systems.*, Boston, Massachusetts, USA, 285-291.

Shipman, F. M. and McCall, R. J. (1997). "**Integrating Different Perspectives on Design Rationale: Supporting the Emergence of Design Rational from Design Communication.**" *Artificial Intelligence in Engineering Design, Analysis, and Manufacturing (AIEDAM)*, 11(2): 141-154.

Singhal, A., Salton, G., Mitra, M. and Buckley, C. (1996). "**Document lenght normalization.**" *Information Processing and Management*, 32(5): 619-633.

Soltysiak, S. J. and Crabtree, I. B. (1998). "**Automatic Learning of User Profiles - towards the personalization of agent services.**" *British Telecom Technology Journal.*, 16(3): 110-117.

Sorensen, H. and Mc Elligott, M. (1995). **An Online News Agent**. *BCS Intelligent Agents Workshop, British Computer Society*, Britain

Sorensen, H., O' Riordan, A. and O' Riordan, C. (1997). "**Profiling with the INFOrmer Text Filtering Agent.**" *Journal of Universal Computer Science*, 3(8): 988-1006.

Sveiby, K. E. (1997). **The New Organizational Wealth: Managing & Measuring Knowledge-Based Assets.**, Berrett-Koehler Publications.

Tauritz, D. R., Kok, J. N. and Sprinkhuizen-Kuyper, I. G. (1999). "**Adaptive Information Filtering using Evolutionary Computation.**" .

Wiener, E., Pedersen, J. O. and Weigend, A. S. (1995). **A Neural Network approach to Topic Spotting.** *Proceeding of the Fourth Annual Symposium on Document Analysis and Information Retrieval (SDAIR'95).*

Yang, Y. and Pedersen, J. O. (1997). **A Comparative Study on Feature Selection in Text Categorization.** *Proceedings of the Fourteenth International Conference on Machine Learning (ICML '97)*

Yu, C. T., Lam, K. and Salton, G. (1982). "**Term Weighting in Information Retrieval. Using the Term Precision Model.**" *Journal of the ACM,* 29(1): 152-170.