CHAPTER 14: DISTRIBUTED INFORMATION SYSTEM

1. **Distributed Information system**
   Definition: *Distributed Information Systems* (DIS) are collections of electronic networked *information resources* (e.g. databases) in some kind of interaction with communities of users; examples of such systems are: the Internet, the World Wide Web, corporate intranets, databases, library information retrieval systems, etc.
   DIS serves large and diverse communities of users by providing access to a large set of heterogeneous electronic information resources.

2. **Automated Data Processing System**
   ADAPS is a collection of software that takes the processing tasks of the water distribution system
   ADAPS programs are used to compute water-data records on an electronic computer.
   The computer operates in parallel with manual data processing system
   The sequence of data processing must be very well established. However there is some flexibility in the sequences so as to improve efficiency, if needed.
   In addition to data processing, ADAPS have additional functions including: initialisation, maintenance, security, backup, recovery, and restart.
3. **File maintenance**
   Story of Credit Suisse – First Boston.
   Story of Merrill Lynch

4. **TalkMine**
   Developed by Computer Research and Applications Group
   Los Alamos National Laboratory, Massachusetts

Talkmine managed DIS is itself a distributed database or search-engine which is continually adapting to and learning from its uses and their patterns of information usage. Each node of the network learns to relate its information to other nodes from repeated interaction with users. With this permanent feedback loop between user and network, the structure of the DIS reflects the knowledge of its community of users. DIS with *Talkmine* function both as collaborative information networks and as ever present, distributed, search engines.

The recommendation system of *Talkmine* is both a content-based and collaborative *(i.e.* used to re-combine knowledge as well as adapt it to users.)

Information Retrieval (IR)

Information Retrieval (IR) refers to all the methods and processes for searching relevant information out of information systems (isolated or part of DIS) that contain extremely large numbers of documents. As the complexity and size of both user communities and information resources grows, the fundamental limitations of traditional information retrieval systems have become evident in modern DIS.

Traditional IR systems are based solely on *keywords* that index (semantically characterize) documents and a query language to retrieve documents from centralized databases according to these keywords - users need to know how to "pull" relevant information from passive databases. This setup leads to a number of flaws (Rocha and Bollen, 2000), which prevent traditional IR processes in DIS to achieve any kind of interesting coupling with users. The human-machine interaction observed in these systems is particularly rigid: Most cannot pro-actively "push" relevant information to its users about related topics that they may be unaware of, there is typically no mechanism to exchange knowledge, or crossover of relevant information among users and information resources, and there is no mechanism to recombine knowledge in different information resources to infer new linguistic categories of keywords used by evolving communities of users. *In other words, traditional IR keeps DIS as static, passive, and isolated repositories of data; no interesting human-machine co-evolution of knowledge or learning is achieved.*

The limitations of traditional IR and DIS are even more dramatic when contrasted with biological distributed systems such as immune, neural, insect, and social networks. Biological networks function largely in a distributed manner, without recourse to central controllers, while achieving tremendous ability to respond in concerted ways to different environmental necessities. In particular, they are typically endowed with the ability to elicit appropriate responses to specific demands, to transfer and process relevant information across the network, and to adapt to a changing environment by creating new behaviors (often from recombination of existing ones). These abilities are precisely what have been lacking in IR.

The ultimate goal of IR is to produce or recommend relevant information to users. It seems obvious that the foundation of any useful recommendation should be first and foremost based on the identification of users and subject matter. In this sense, the goal of recommendation systems can be seen as similar to that of most biological systems, in particular immune systems: to recognize agents (users) and elicit appropriate responses from components of the distributed information network. Furthermore, the information network should learn and adapt to the community of agents (users) it interacts with - its environment.

1. A means to recognize ***users***.
2. A means to characterize ***information resources***.
3. A 2-way means to exchange knowledge between users and information resources: a ***conversation*** process. As information resources become more and more complex, we cannot expect a simple 1-way query ("pull") to work well. Instead, we need a means to integrate the interests of the user with the knowledge specific to each information resource via an interactive recommendation process ("push").
4. ***Adaptation*** mechanisms. We also want DIS to adapt to their community of users, as well as to exchange and re-combine knowledge leading to evolvability and creativity.

Recommendation systems are typically based on human-machine interaction mediated by intelligent agents, or other decentralized components, and come in several varieties:

1. In *content-based* recommendation, user profiles are created based on the system's keywords. Documents are recommended to users according to the similarity of their profiles and the similarity of keywords constructed from a semantic distance function obtained from the associations between keywords and documents. Two documents are close when they are classified by many of the same keywords. This is the case of systems such as *InfoFinder* (Krulwich and Burkey, 1996), *NewsWeeder* (Lang, 1995), and many systems developed for the routing task at the TREC Conferences (Harman, 1994).
2. In *collaborative* recommendation no description of the semantics or content of documents is involved, rather recommendations are issued according to a comparison of the profiles of several users that tend to access the same documents. The comparison depends on a distance function between user profiles, defined not by keywords, but on the sets of actual documents retrieved. Two user profiles are close when their users have retrieved many of the same documents. This is the case of systems such as *GroupLens* (Resnick et al, 1994; Kostan et al, 1997), *Bellcore Video Recommender* (Hill et al, 1995), *Ringo* (Shardanad and Maes, 1995). When user feedback is allowed, this type of recommendation is known as *Information Filtering* (Good et al, 1999). For a description of the collaborative recommendation framework see Herlocker et al (1999).
3. In *structural* recommendation, data-mining techniques are employed on the relations among documents and keywords, to discover related documents or documents of particular importance

(authorities) in a given information resource. A large portion of work in this area, is concerned with the analysis of the graph structure of Web Hyperlinks (regardless of document keywords), e.g. work pursued under the *CLEVER* Project (Kleinberg, 1998; Chakrabarti et al, 1999), or other graph-theoretic approaches such as Watts' (1999) Small World graphs. A second large area of research is concerned with the semantic relations between documents and keywords, which are analyzed with algebraic techniques such as Singular Value Decomposition, known in IR as Latent Semantic Indexing (LSI) (Berry et al, 1994; Kannan and Vempala, 1999). *Documents are recommended to users according to the way they are associated with other documents and/or keywords: the semantic structure of information resources.*

4. In *collective* recommendation, the behavior of communities of users is integrated, and utilized to adapt the structure (the pattern of associations) of information resources. This kind of system tracks the paths users follow in the structure of information resources as they retrieve documents. The more certain sets of documents tend to be retrieved together in paths followed by different users, the closer they become in the structure of the information resource. This type of algorithm employs the distributed behavior of a collection of users to adapt DIS, resulting in systems that learn the interests of their communities of users much in the same way as social insects discover paths based on the pheromone trails left behind by other insects in their colony (Rocha and Bollen, 2000), thus, in time, recommending more and more appropriate documents. This is the case of Adaptive Hypertext systems (Brusilovsky et al, 1998; Bollen and Heylighen, 1998; Eklund, 1998), Knowledge Self-Organization (Johnson et al, 1998; Heylighen, 1999), as well as the work on the collective discovery of linguistic categories (Rocha, 1997a, 2000) detailed below.

Content-based systems depend on single user profiles, and thus cannot effectively recommend documents about previously unrequested content to a specific user. That is, these systems cannot compare and recommend related documents characterized by keywords not previously collected into a given user's profile. Conversely, pure collaborative systems, match only the profiles of users that (to a great extent) have requested exactly the same documents; for instance, different book editions or movie review web sites from different news organizations may be considered distinct documents.

The shortcoming of structural approaches is that they assume that the existing, often static, structure of an information resource contains all the relevant knowledge to be discovered. However, it is often the case that such structure is very poorly designed. On the web in particular, the hypertext links are often not created between important documents, due perhaps to the hurried way in which web sites are created. Indeed, the Web is often more a repository of isolated documents, than a good example of a hypertext fabric. The same applies to the keyword/document relations necessary for LSI.

Collective approaches have the important advantage of adapting to the collective behavior of users, even as it develops in time. This way, a poor initial structure can improve, by creating, strengthening or weakening associations among documents or between documents and keywords. Furthermore, collective recommendation systems can operate without storing individual profiles, thus offering a more private platform for recommendation. Indeed, recommendations are issued according to the adapted structure of the information resources, not according to user profiles. Users can be seen as anonymous social agents. Furthermore, as we shall discuss later, the adapted information resources allow us to capture the knowledge traded by a community of agents. *Nonetheless, a disadvantage of collective approaches is that they implement a positive feedback with their communities of users, possibly leading to an excessive adaptation to the interests of a majority of users, thus reducing the diversity of knowledge by recommending only the most retrieved documents in a given area: e.g. the "best of" lists found at Web sites such as* Amazon.com *- this is the so-called "curse of averages".*

It is clear that good recommendation systems require aspects of all approaches to avoid the shortcomings of each individual one. This is the case, for instance, of *Fab* (Balabanovi and Shoham, 1997) and *Amalthaea* (Moukas and Maes, 1998), which are both content and collaborative recommendation systems. *This way they can discover similar users who have not simply retrieved many of the same exact documents, but documents characterized by many of the same keywords. Furthermore, keywords from documents that users have not actually retrieved, may be added to their profiles because they belong to the profiles of other similar users.*

Still, neither *Fab* nor *Amalthaea* (nor similar systems) adapt the structure of their information resources with collective user behavior, nor do they use the data-mining techniques of structural algorithms to characterize the knowledge those store. In this sense, they cannot capture the evolving nature of the knowledge of communities of users. In other words, even though they are able to characterize the interests of individual users (both with documents and keywords), the structure of information resources (e.g. Web hyperlink structure or document/keyword matrix)

remains unchanged. *Furthermore, they rely on individual user profiles, and there is also not an explicit means to discover the knowledge categories that particular communities of users employ.*