A Proposal for an Experimental Platform on Collaborative Information Retrieval

Sergio Cleger-Tamayo  
Departamento de Informática.  
Universidad de Holguín, Cuba  
E-mail: sergio@facinf.uho.edu.cu

Juan M. Fernández-Luna, Juan F. Huete  
Departamento de Ciencias de la Computación e I.A.  
Universidad de Granada, Spain  
E-mails: {jmfluna,jhg}@decsai.ugr.es

Ramiro Pérez-Vázquez  
Departamento de Ciencias de la Computación  
Universidad de Holguín, Cuba  
E-mail: rperez@uclv.edu.cu

Julio C. Rodríguez-Cano  
Departamento de Informática  
Universidad Central de las Villas, Cuba  
jcrodriguez@facinf.uho.edu.cu

ABSTRACT

Current research on Collaborative Information Retrieval area reveals a significant synergy among Information Retrieval, Human-Computer Interaction and Computer-Supported Cooperative Work disciplines. This fact has a high incidence in the modernization of Information Retrieval Systems. In this paper, we describe both our current research efforts to design an experimental platform for collaborative information retrieval, and our progress towards implementing it as a pluggable groupware. The platform’s core has been designed as a framework, to maximize its reusability and adaptability to experiment with collaborative information retrieval techniques, with a minimal programming effort.

KEYWORDS: Collaborative Information Retrieval, Collaborative search, Computer-Supported Cooperative Work, Human-Computer Interaction, Personalization, Recommendation.

1. INTRODUCTION

Nowadays, as the amount of information available causes information overloading, the demand for efficient approaches to better information access has increased. In addition, with the advent of Internet participation era, where social navigation systems (e.g. social networking sites) encompass a variety of techniques that help people work together to help each other [1], a new problem in Information Retrieval (IR) discipline has begun to gain attention, mainly because IR researchers considered the IR process as a one-person activity, or have only considered the asyn-
chronous collaboration as in the Recommendation Systems [2, 3, 4, 5]. These determinations prevent users with shared information needs from collaborating properly.

In the last few years, some researchers have realized that collaboration is an important feature, which should be analyzed in detail in order to be integrated with professional IR systems, upgrading them to Collaborative Information Retrieval (CIR) systems [6, 7, 8, 9, 10, 11].

CIR-based technologies may change our life while getting better communication with people who have the same information needs that we have. In this way several disciplines contribute to CIR. Some researchers highlight three main disciplines [8, 12, 13]: IR, Human-Computer Interaction (HCI) and Computer-Supported Cooperative Work (CSCW), as shown in Figure 1. Other authors have only focused on IR and HCI disciplines, in order to create a new kind of search systems within Human-Computer Information Retrieval (HCIR) area, which depends on continuous human control during the searching process [14].

Some CIR prototypes and systems have been developed, but most of these groupwares are not well designed to support collaboration (except a few of them [12]), because the experience gained from programming traditional IR systems cannot be simply extrapolated to CIR systems. Consequently CIR area involves a number of new issues:

1. How searching tasks can be usefully divided among many users.
2. How a user can exploit the knowledge generated by another one.
3. How to facilitate efficiently the groups’ awareness of searchers.

There have been some previous research efforts toward enabling collaboration when searching. However, most of these prior systems are designed for specific problem or devices, rather than general-purpose CIR. For example, Fischlár-DiamondTouch [15], uses a conventional backend search engine for text and image/keyframe searching, and a novel multi-user collaborative tabletop interface to support collaborative group searching. SearchTogether [9] is a prototype which enables remote users to synchronously or asynchronously collaborate when searching the Web. It supports collaboration with several mechanisms of awareness, division of labor, and persistence. CoSearch [16] is a system to improve the experience of co-located collaborative Web search, by leveraging readily available devices such as mobile phones and extra peripherals. Cerchiamo [17] is a collaborative exploratory searching system that allows teams of searchers to explore document collections synchronously. Working with Cerchiamo, team members use independent interfaces to run queries, browse results, and make relevance judgments. The system mediates the team members’ searching activity, by providing and reordering search results and suggested query terms, based on the team’s actions. MediaMagic [18] is an interactive video search system. It allows the searcher to quickly assess query results and easily pivot those results off to form new queries. Using MediaMagic, two or more users with common information needs search together, simultaneously. The collaborative system provides tools, user interfaces and, most importantly, algorithmically-mediated retrieval to focus, enhance and augment the team’s search and communication activities.

We present in this paper our current research efforts to design CIRLAB (Collaborative Information Retrieval Laboratory) as an experimental platform in CIR area. CIRLAB allows researchers to use a set of user interfaces, and to experiment CIR techniques merged with personalization techniques, search engines and collections with a minimal programming effort. Unlike other CIR tools, CIRLAB is not designed for specialized problems. It presents the following features:

- Easily experimentation in the CIR context.
- Recognizing user groups with shared information need.
- Supported by recommendation algorithms and techniques, speeding up access to information and user coordination.
- Recommendation of items to users or groups of users.

One of goals of this paper is to describe the basics of CIR area, which we presented in Section, and its relation with personalization techniques, appearing indistinctively in sections and. The Section contains the general description of the experimental platform proposed. We dedicate Section and Section to introduce our main ideas about the architecture design and our progress towards of CIRLAB implementation and instantiation. Finally, we conclude this paper in Section by summarizing the exposed topics.

2. OVERVIEW

Currently there are two viewpoints on how to use IR systems. First, a traditional IR system aims to satisfy a user’s information need [19, 20]. Second, a modern IR system supports multiple users interacting simultaneously in the search process [21]. This particular kind of groupware (Subsection) is within CIR area. Up to this day, there is no
agreement about the terms and concepts in CIR. P. Hansen and K. Järvelin [8], two of the first researchers who considered collaboration as an important component in IR process, defined CIR as “an information access activity related to specific problem solving activity that, implicitly or explicitly, involves human beings interacting with other human(s) directly and/or through texts (e.g., documents, notes, figures) as information sources in a work task related information seeking and retrieval process either in a specific workplace setting or in a more open community or environment”.

Let us enumerate some basic goals of CIR systems where we expect tagging to be helpful:

1. Reduction of searching time among users with shared information needs.
2. Quality improvement of professional IR systems.
3. Easy access to, or interaction with, experts on a specific topic.

As from this last goal, some authors have proposed different kinds of CIR systems, for example in [22] J. Pickens and G. Golovchinsky stated two types of search groups, a) domain experts search together, and b) domain experts working with search experts.

2.1. Different Perspectives in CIR

Salton et al. [23] describe IR as a cross disciplinary field, wherein the main question is “how does one find the relevant documents in a collection of documents given a user query?”. Today, hundreds of millions of people engage in IR every day when they use a web search engine or search their email. IR is quickly becoming the dominant form of information access, overtaking traditional database style searching [24].

HCl and IR disciplines have both developed innovative techniques to address the challenge of navigating the complex information spaces, but so far their insights often failed to cross disciplinary borders. Human-Computer Information Retrieval (HCIR) has emerged in academic research and industry practice as the study of IR techniques that bring human intelligence into the searching process. This field brings together research in the fields of IR and HCI in order to create new kinds of searching systems that depend on continuous human control of the searching process [14]. In addition CSCW can be blended with HCIR in order to show CIR as a new interesting research area. CSCW is defined in [25, 26] as an activity carried out by a group of persons, but coordinated and assisted by computers. The term collaboration (which is not only included in the definition, but also results important to analyze from CIR perspective), is defined as the interaction between humans implicated in a joint activity, even though currently there is no consensus about the exact significance of the terms collaboration and cooperation, since they are used by some authors in topics related to CSCW. Other authors described different approaches to this concept: Collaboration is a superior activity compared to cooperation [27]. Whenever people cooperate they share the same goals but act independently, or on the same task but on different parts. On the other hand, when they collaborate, even though they cooperate, they also work altogether in related tasks towards the same ending result. Collaboration provides two important characteristics: something new is created and it is created within a space of common representation.

When CSCW is centered in small groups of persons sharing a common goal and needing to communicate, it is considered cooperation. If we find ourselves in a large organizational system, where there are objectives frequently in conflict, then it is considered collaboration [28]. The meaning of these terms is defined by their field of use: the term cooperation comes from the HCI or telecommunications and software companies; and the term collaboration is commonly used in the areas of information systems, universities and governmental institutes. Collaboration has also been considered as a synonym of cooperation [29].

Another relevant issue in CSCW is the use of the term groupware as a reference to the multiuser software that supports collaborative systems, even though it is commented that with CSCW other technologies besides computers may be used, that it may support the competition as well as the coordination, and that it can also assist in casual and social interactions not related to those typical of the work area. One of the most general definitions is [30] “Groupware is the software and hardware for shared interactive environments.”

2.2. Features of CIR Systems

Most of the CIR systems developed recently, include some common features: session persistence, division of labor, knowledge sharing and awareness. In this section we describe each one and make reference to the groupware in which they are involved.

2.3. Session Persistence

Storing a search session in a persistent format is a key requirement to facilitate collaboration during the session, revising the search at a later time, or sharing the results of a search with others [31].
2.4. Division of Labor

Morris’s survey in [31] describe ad-hoc methods to avoid duplication of effort during a searching task, such as dividing up the space of potential keywords, searching engines, or sub-tasks among different group members. Supporting mechanisms for dividing up and sharing work among participants is important to the success of a UI for multi-user search.

2.5. Knowledge Sharing

In any collaborative setting there will be a large and diverse knowledge base shared among group members. Each one will bring their own experiences, expertise and topic knowledge to a particular searching task. What is needed is a way to enable the sharing of knowledge within the group [21].

2.6. Awareness

Another key requirement for collaborative search is a mean to provide awareness of the collaborators activities. This kind of information is useful in both synchronous and asynchronous settings [31].

2.7. Recommender Systems

The Recommender System (RS) field is defined as the area of Artificial intelligence [32, 33] where researchers concentrate explicitly in evaluating the relevance of those elements that the user does not know. Two main categories stand out, content-based RS, in which recommendations are made based on the users previous selections, and the collaborative RS, in which recommendations are based on selection of users with similar tastes and preferences. In these two mentioned approaches, data mining algorithms are used to determine the usefulness of existing relations, conceiving a collective intelligence to show information adapted to the client preferences, information that in many cases was not expected to be found and may result useful.

Currently, the interest for recommendation systems continues to grow, making integral parts of some important sites of e-commerce such as Amazon.com [34], where recommendations for books, DVDs and other products are made, and in which Jeff Bezos claimed to excellence: “Each one of our clients will have a shop to his taste”. Also we have MovieLens.com where movie recommendations are made, or Adaptiveinfo.com, a news recommendation system.

A high number of RSs work based on Collaborative Filtering (CF) views, being understood as the method to make automatic predictions (filtering) with respect to a users interest, based on the information in existence of various users (collaborators). The collaborative filter is like a “black box”, something uncontrollable in which we ought to trust.

CF systems, also known as social filtering, depend on a data base of items and users, and recommend objects to an objective (active) user, based on the opinions of other users with similar taste. This means that the principal aim of RSs is to predict the usefulness of objects to a particular user, based on a users evaluation data base. The main idea is the automation of the ”word of mouth” process through which persons recommend products or services to each other.

In general terms, RSs can be a reality in any business or system that has many options or opinions to help users decide (decision making). For example, the television industry, professional search, search for similar persons, or news services.

Nonetheless, little has been the use of RSs in CIR systems. For example, Morris [13] proposes three techniques that can enhance the value of collaborative search tools using personalization: groupization, smart splitting, and group hit-highlighting.

3. PLATFORM DESCRIPTION

Considering the harsh conditions CIR researchers have to face, in order to create new user interfaces, algorithms and evaluation metrics, we are designing a platform which allows flexible experimentations with CIR and personalization techniques. It is based on a Service Oriented Architecture (SOA) to perform rapid and low-cost pluggable groupware development and improving overall system quality.

We did not select the Web scenario because it involves problems we had to consider. The Web is a good environment to provide communication and asynchronous collaboration mechanisms, but it is not simple to implement synchronous collaboration on it, specially when processes require intense users interaction.

Groupware are often classified along two axes: whether they support distributed or co-located users, and whether these users collaborate in a synchronous or asynchronous fashion. We are designing an Internet Communication Engine (ICE)-based on a framework for our experimental platform which supports remote collaborators, working either synchronously or asynchronously.

ICE is an object-oriented middleware platform. Basically, this means that ICE provides tools, APIs, and library sup-
port for object-oriented client-server applications development. ICE applications are suitable for use in heterogeneous environments: client and server can be written in different programming languages, can run on different operating systems and hardware architectures, and can communicate using a variety of networking technologies. The source code for these applications is portable regardless of the deployment environment [35].

A framework is a pattern arising at the system architectural level. It is a collection of abstract classes that provides an infrastructure common to a family of applications. It dictates certain roles and responsibilities amongst its classes, and specifies the standard protocols for their collaboration.

Our proposal besides contemplating various of the ideas in recent CIR investigations, incorporates personalization techniques which allow a better integration between the different stages of the IR process amongst various users with shared information needs.

In the modern Web, as the amount of information available causes information overloading, the demand for personalized approaches to information access increases. Personalized systems address the overload problem by building, managing, and representing information customized for individual or multiple users. This customization may take the form of filtering out irrelevant information and/or identifying additional information of likely interest for the user or group of users [2, 3].

CF plays a major role when the sources for the search processes, imply not only documental search (called documents), but also other types of sources (items such as images, videos, audio or composed items) and are not analyzable by simple techniques like those used by traditional searching engines. That is why those who collaborate, pay tribute to the previous searching processes. CF uses the assumption that people with similar tastes will rate things similarly [3].

Division of Labor was described in subsection as one of the most important characteristics of CIR systems. Through personalization techniques an efficient division of labor can be done, on the basis of user profiles and the persistence of the collaborative searching sessions [2].

4. ARCHITECTURE DESIGN

This section describes the architecture of CIRLAB (Collaborative Information Retrieval Laboratory). CIRLAB has five main tiers: user, server, information, persistence and evaluation (Figure2). All of these tiers are configurable with the purpose of having a flexible platform for experimentation with CIR techniques.

4.1. User Tier

The user tier consists of a set of client’s views that we show in Figure2. The individual client view is similar to the classical IR system front-end (e.g. Google, Yahoo! or Windows Live Search in the World Wide Web). It allows showing the user interface of the collaborative searching portal, in which the user may create a searching session or join to one created explicitly.

A collaborative searching session wraps a group of users with shared information needs. When a user join to a collaborative searching session, his individual client view is updated with CIR features (Subsection) and converted into a distributed client view.

A user can easily pass from an individual view (Figure2) to a distributed view moving through three stages during the transition. The first stage is group association, which can be done in an explicit or implicit manner. Various users can agree to make a search, for example, by being a research team at a school. Another way in which various users can associate to make searches is based on recommendations made by a search agent based on their profiles (Figure3). Once various users are explicitly associated they can create a collaborative searching session (CSS) and get to the third stage where they can also be making searches from the distributed view. It can also occur that a user joins a CSS already created from recommendations made by the agent.

All of user’s views are represented in CIRLAB as a Java hybrid-application, this means users may interact with in-
terfaces on desktop (as a Java Swing application) or in browsers (as a Java Applet). The user view components can communicate with server tier through the ICE-based framework. We select a Java hybrid-application because it is more efficient in order to interchange with the server via TCP/IP than others client technologies via HTTP.

4.2. Server Tier

In this tier various related layers co-exist. One of the layers consists of classic search engines like Apache Lucene, Lemur, Terrier or Google Search API, which can be used indistinctively to search in the information tier, according to configuration. In relation to the items returned by the engines, different filtering techniques may be applied to recommend better items from the whole set. Most of the CF systems we have discussed so far have been systems which use the group as a whole to help each individual user.

The results of classic search engines can be used as inputs for the layer of CIR techniques. Moreover, these inputs can be stored in persistence tier and then apply some algorithms like P2R, which applies a hit-matrix to the retrieved pages of the conventional Web search engines to refine the search results. The hit-matrix keeps track of the frequently clicked pages for each frequent phrase of the queries [36].

The personalization layer is a very important one because it allows the use of techniques as Groupization, Smart Splitting and Group Hit-Highlighting, proposed by Morris in [13]. These three techniques can enhance collaborative searching tools, and demonstrate their value using empirical data.

We may also remark that many CIRLAB elements are easily managed, thanks to Configuration Tier. From this tier a researcher may specify which search engines and algorithms shall be used, which will be the collections, what information will become persistent, and so on. This way researchers will be able to make previous plans, simplifying not only CIRLAB adaptation to diverse environments and particular problems, but also evaluate different search scenarios.

4.3. Information Tier

The information tier represents the data collections that could be used from the platform. This consists basically of allowing the use of data collections or Internet to test the available techniques in the server tier.

4.4. Persistence Tier

This tier allows making persistent all data that results significant for CIR techniques as well as personalized data. Previous searches should be evaluated within a range of values for future filtering tasks and recommendations, these evaluations could be either explicit or implicit. Many personalization techniques are based on, or include, some type of user profile, a data instance of a user model which is applied to adaptive interactive systems. User profiles may include demographic information, e.g., name, age, country, education level, etc., and may also represent the interests or preferences of either a group of users or a single person. In order to construct an individual user profile, information may be collected explicitly, through direct user intervention, or implicitly, through agents that monitor user activity.

We may also remark that many CIRLAB elements are easily managed, thanks to Configuration Tier. From this tier a researcher may specify which search engines and algorithms shall be used, which will be the collections, what information will become persistent, and so on. This way researchers will be able to make previous plans, simplifying not only CIRLAB adaptation to diverse environments and particular problems, but also evaluate different search scenarios.

4.4. Evaluation Tier

Golovchinsky et al. remark in [12] that the evaluation is another challenge in CIR, because traditional information retrieval systems are usually evaluated based on recall and precision measures, and these metrics assume a single logical searcher (even if more than one person contributed to the final search results). Therefore, it is necessary to develop new metrics to assess the contributions of multiple team members to collaborative search processes.

In [11], Foley et al. presents a measure to capture the quality and diversity across collaborating users’ ranked lists. Those authors propose to measure the total number of unique relevant documents across user’s ranked lists at a certain cutoff and use this figure as group score. This idea can be applied in the Evaluation Tier, in addition to others evaluation measures that we may design.
5. AN APPROACH TO IMPLEMENTATION AND INSTANTIATION

For the implementation of CIRLAB we use the framework FreeTribe (FRamework for dEvElopment of disTRIButed groupwarE). This framework is a proposal for distributed groupwares development, made by the Distributed Systems Research Group, from Information Technologies Department in Universidad de Holguín.

FreeTribe uses ICE as middleware, Java as programming language and is supported by AMENITIES (A MEthodology for aNalysis and desIgn of cooperaTIve systEmS) methodology [37], to ensure an adequate reutilization of both design and source code. AMENITIES is a methodology which allows addressing the analysis and design of CSCW systems systematically and which facilitates subsequent software development. It allows the realization of a conceptual model of cooperative systems and focuses on the group concept. It covers significant aspects of both group behavior (dynamism, synchronization, etc.) and structure (organization, laws, etc.). The resulting specification contains relevant information (cooperative tasks, domain elements, person-computer and person-person dialogues, etc.) to the creation of the user interface[38].

To deliver recommendations, we select Taste class library. Taste is a flexible, fast collaborative filtering engine for Java. The engine takes users’ preferences for items ("tastes") and returns estimated preferences for other items. Taste is designed to be enterprise-ready; it is designed for performance, scalability and flexibility [39].

An important stage during generic solutions testing, is their instantiation in test cases. CIRLAB must evolve until it fulfills all requirements for a CIR system, also taking advantage from new features added to FreeTribe because, at this stage, instantiating CIRLAB also requires direct interaction with FreeTribe and ICE (Figure 4).

CIR systems and prototypes proposed up to this day, were applied to several domains, such as travel planning, organizing social events, working on a homework assignment[9], in medical environments [40, 41], or multimedia contexts [9, 18].

For pragmatic CIRLAB validation purposes we identified software development as applicability field, where many evidences of programmers collaboration over development process can be found [42, 43]. We could also mention, as a feature to be added to IDEs, the collaborative source code search. Currently there are some source code searchers (e.g. Google Code Search, Krugle, CodeFetch, Koders, Codase), some of them such as Koders, can be integrated to IDEs, but all those interfaces are one-person oriented.

To instantiate CIRLAB as a collaborative source code searcher, the individual view interface could be used in first place, followed by distributed view so the potential benefits of CIR systems could be evaluated. In Server Tier some personalization models could be used, always regarding the developers preferences. Afterwards, a plan could be made to determine which collections shall be used (source code repositories and books on programming), which search engines and algorithms shall be used to perform tests.

As final step, Evaluation Tier could be used to confirm the existence of significant differences between two approaches: one, source code searchings performed by a group of programmers using the individual view and collaborating through conventional means such as classic search engines, e-mail or chat; and two, another group of programmers who used the distributed view and ad-hoc CIR algorithms and search engines.

6. CONCLUSIONS

The CIR community has been developing some groupware in the last few years, but many of them address only specific problems or do not adequately support collaborative interaction. Therefore is convenient to have an experimental platform that allows to test new algorithms which use inputs from multiple users and compares its searching results with a single user activity. In that sense we propose CIRLAB as a reusable and adaptable tool which allows an easy experimentation with CIR techniques.

Session persistence, division of labor, knowledge sharing, and awareness shows up within CIR as important elements to have in consideration for its notable benefits to users with shared information needs. Some of these elements can mix up with personalization techniques in order to improve the group search interaction.

Figure 4. CIRLAB Instantiation
There are many contexts where pushing IR horizons towards CIR would be a vantage. Some CIR systems and prototypes on Web search, medical environment and multimedia, have been proposed. This paper proposes the software development as another study area for CIR systems to deploy. Software development is a collaborative process, where teams of developers work together to design, solve problems, and particularly, share their source code search results, and also send by e-mail or chat source code fragments to their colleagues or an URL where doubts may be cleared.

ACKNOWLEDGEMENTS

This work has been jointly supported by the Spanish Ministerio de Ciencia e Innovación, under project TIN2008-06566-C04-01, and the Consejería de Innovación, Ciencia y Empresa de la Junta de Andalucía, under project TIC-276.

REFERENCES


