This paper reports an exploratory study on resources and resource sharing among researchers in Intelligent Information Access (IIA). The investigation consists of two stages. In Stage One, the authors conducted a content analysis to identify resources used in 145 research papers and reports in two subfields of IIA; and in Stage Two, the authors carried out an online survey of IIA researchers to understand resource-sharing channels and the researchers’ perspectives on resource sharing. The results demonstrate that IIA researchers make use of various types of resources developed by others. Most of these resources are knowledge sources or software systems that are freely available online. However, IIA researchers encounter various difficulties during the course of resource acquisition and use. The study suggests that a resource management system built on a well-established knowledge-management model could greatly facilitate the creation, sharing, and use of resources in the IIA community.

1. Introduction

Web search engines play an important role in the modern information society. They accept Web users’ keywords and return a list of ranked Web pages. However, they are neither efficient nor effective when the users’ information needs go beyond a list of returned Web pages. For example, a Web user may just want to find out the exact answers to such questions as “what diseases are prions associated with” and “who is the lead singer/musician in Nirvana?” Or, an American investor may want to find publications related to a foreign stock, but many of the publications furnished are written in a language other than English.

Intelligent Information Access (IIA) technologies can satisfy the above information needs. IIA technologies make use of human knowledge or human-like intelligence to meet special inquiries of information users. The problems presented in the two examples above can be addressed by Automatic Question Answering (QA) and Cross-Language Information Retrieval (CLIR) systems. Specifically, a QA system can return answers or facts instead of a list of Web pages, and a CLIR system can find relevant documents or Web pages in languages different from users’ queries. In addition to QA and CLIR, IIA contains many
other subfields such as Automatic Classification and Clustering, Text Summarization, and Information Extraction (Chen, Li, & Xuan, 2006). IIA technologies have become popular research topics because of their substantive potential applications in many fields such as corporate online systems and portals (Chung, 2008), e-commerce (Benetti, et al., 2002), and digital libraries (Chen, Li, & Li, 2006). Researchers and practitioners from a large variety of disciplines have become interested in IIA research or in applying IIA technologies for research and educational purposes.

IIA research typically makes use of various resources in order to develop advanced IIA systems to achieve effective information access (Kishida, et al., 2004). For example, Chen and colleagues (2005) applied about 10 resources to develop an experimental Chinese-English CLQA system. For many researchers, the needed resources are not always available. Here resources are broadly defined as any knowledge sources such as ontologies, annotated corpora, test collections, and various types of software such as information retrieval (IR) systems or search engines, natural language processing (NLP) tools, classification systems, and/or machine translation systems. Moreover, developing these resources usually proves to be expensive and labor intensive. For instance, the widely-used English lexicon WordNet took years to be developed with approximately three million dollars of funding mainly from government agencies (Wikipedia: http://en.wikipedia.org/wiki/WordNet). Consequently, IIA researchers either look for freely available resources or collaborate with each other on resource development. Resource sharing has become essential for the advancement and application of IIA technologies.

Little research has examined the status of resource sharing among IIA researchers. Also lacking is an effective model of resource management that could establish a foundation for effective administration of the resources developed by the IIA community and other related communities such as the lexical resource development group. It is unclear what the most useful resources are, how IIA researchers obtain the resources they need, what the resource-sharing channels are, and whether their quests for resources have been satisfied through current resource-sharing channels.

We designed this exploratory study to understand the current resource-sharing status among IIA researchers as the first step towards developing an effective resource management model for IIA. The investigation consists of two stages: (1) A content analysis of publications in two subfields of IIA, CLIR and QA, to identify the types of resources researchers use and the characteristics of these resources; and (2) a survey of IIA researchers to understand their approaches to and perspectives on resource sharing and acquisition. The two stages are integrated in order to obtain a complete picture of the current resource-sharing statuses among IIA researchers. Additionally, investigation in the first stage facilitates the design of the survey instruments for the second stage, and the results from the second stage justify the findings from the first stage.
This paper is organized as follows. The next section provides an overview of related work in knowledge sharing and IIA resource sharing and acquisition. The following section describes the study’s research design. Afterwards, we discuss the results and findings from the content analysis and the survey. The paper concludes with discussions on the issues related to resource sharing and acquisition and calls for future research to improve resource sharing and management of IIA.

2. Resource Sharing and Management in IIA

This study is guided by the theories of knowledge sharing and knowledge management. Knowledge is “justified true belief” as defined by Plato (Small & Sage, 2005) and stands closely related to information - useful data or objects that could impart knowledge (Buckland, 1991). In reality, knowledge and information are terms habitually interchanged in information science and decision-making science.

Knowledge management (KM) is a term widely used but not well defined. This paper adheres to the forms of knowledge presented by Boersma and Stegwee (1996) and the results of a study conducted by Davenport, De Long, and Beers (1998). Boersma and Stegwee identified four forms of knowledge that should be managed: human, documented, mechanized, and automated, in which documented knowledge is knowledge stored in books or documents and automated knowledge consists of software systems that can perform specific tasks. IIA resources are either documented or automated knowledge. Davenport and colleagues surveyed 31 knowledge-management projects and summarized their major objectives: “create knowledge repositories, improve knowledge access, enhance knowledge environment, and manage knowledge as an asset” (p. 44). Knowledge management has been explored in mainly organizational settings in which people create and share knowledge for the purposes of improving the performance of the organization. Bukowitz and Williams (1999) present a useful KM process framework addressing both the tactical and strategic knowledge-management processes, which will be discussed later in the conclusion session.

The IIA community is a loosely organized virtual community considered as Community-of-practice (CoP) in which groups of people share common interests and engage in activities to achieve similar goals (Hafeez & Alghatas, 2007). The strategies of KM for IIA can only be determined after a careful analysis of the ways the community generates and shares knowledge contained in the various resources.

Few systematic analyses and summaries of the use of resources have been found despite the influx of more researchers from multiple disciplines. Certain tools and services for automatic classification are presented at http://searchtools.com/info/classifiers-tools.html, with a brief annotation for each resource (Search Tools Consulting, 2003). Amitay (n.d.) has built the Web IR and IE site, which is a collection of online resources for research in
Information Retrieval and Information Extraction. The ACL NLP/CL Universe (http://tangra.si.umich.edu/clair/universe-rk/html/u/db/acl/) is a “publicly extensible indexed compilation of Web resources related to computational linguistics.” It contains links to NLP-related information and subject-specific resources. The system was last updated on August 14, 2006. The Natural Language Software Registry (NLSR) at http://registry.dfki.de/ was an initiative of the ACL (The Association of Computational Linguistics) hosted by the German Research Center for Artificial Intelligence (http://www.dfki.de/web). According to its website, NLSR provides a structured listing and descriptions of available NLP products. By August 2, 2007, 504 records were in the system.

These lists and websites allow other researchers to review each individual resource by following the links set up on the Web pages. Fewer provide (such as NSLR) well-designed submission, search, and browse functions. However, they do not have complete coverage, nor do they provide information on how the resources were actually used.

Our study differs from the above efforts in that our ultimate goal is to provide insights in the systematic management and effective use of existing IIA resources. Our research design is presented below.

3. Research Design

3.1. Research Objectives and Research Questions

This exploratory study seeks to understand resources and resource sharing in IIA. It initiates an investigation of effective resource development and sharing in IIA. The objectives of this research include identifying types of resources that have been used in IIA research and/or development projects, the characteristics of these resources, the channels of resource acquisition and sharing, and the researchers’ perspectives on effectiveness of current resource-sharing channels. Specifically, the following research questions will be investigated:

1. What types of resources are used in IIA research and development?
2. What are the characteristics of these resources?
3. How do IIA researchers acquire and share resources?

3.2. Stage One: The Content Analysis of QA and CLIR Publications

Stage One started in 2005. It is designed to investigate the first two research questions. The content analysis contains four major components: data selection, data reduction, inference, and analysis (Krippendorff, 1980). For data selection, we decided to examine published research papers and reports to locate resources actually used by CLIR and QA researchers. Moreover, we are interested in how the same resources are used by different researchers. As peer-reviewed conference papers and the reports published at the three well-known forums
(TREC, NTCIR, and CLEF) are the most influential for QA and CLIR, we decided to collect latest papers published in the six best-known international conferences and workshops. These conferences and workshops include the 2005 ACM Special Interest Group on Information Retrieval (SIGIR) Annual Conference, 2005 Annual Conference of the ACL, 2005 Human Language Technology Conference (HLT), 2004 TREC, the 2004-2005 NTCIR Workshop, and 2005 CLEF. As a result, 145 CLIR and QA research papers or reports were collected as samples for content analysis. They include 82 papers in QA and 63 papers in CLIR.

The selected research papers are typical scientific documents addressing algorithms, models, or methods for CLIR or QA system development or implementation. These papers normally include a description of the related literature, the algorithm(s) to be examined, the experimental setting, the actual experiments to test the algorithm(s), and the evaluation results. Any resource used during the research is addressed by name in the respective paper. In many cases, the researchers also mention its source, its function, and how the resource is used in the experiment. The purpose of our content analysis is to code the above information in each paper for further data analysis.

The coding is carried out based on a coding scheme developed in a previous study by Chen, Li, and Xuan (2006). The coding scheme, as illustrated in Figure 1, serves as the data language for data recording, analysis, and inference. It contains three top categories: Resource Type, Resource Acquisition, and Usage, labeled as categories A, B, and C, respectively. Generally speaking, this coding scheme is a combination of classification schemes for two different units of analysis. Categories A and B are used to code each individual resource, while Category C is used to code each occurrence of resource in each sample research paper. Category A, Resource Type, categorizes resources according to their content or function. Category B, Resources Acquisition, classifies resources according to their acquisition approaches. And Category C, Usage, examines how resources have been used in each individual IIA research project/experiment from different angles.

3.3. Stage Two: Survey of the IIA Community

Stage Two started in 2006. It employs a Web-based survey aimed at collecting the opinions and perspectives of IIA researchers regarding their channels of resource acquisition and sharing. The original questionnaire contains 17 multiple-choice and 12 open-ended questions, but only the results of questions related to this study are reported due to the limits of paper length.

Participants were selected using the purpose sampling method. Invitations were sent to relevant mailing lists, such as SIGIR-List, Corpora-List, NTCIR 2006, CLEF, and TREC. Anyone who subscribed to one of these lists is qualified to take the survey. It is difficult to estimate the population of IIA researchers as the field is constantly growing. Based on a
rough count of authors of conference proceedings and of participants of relevant tracks at these conferences, active researchers number is around 300 in areas of CLIR, CLQA, QA, classification and clustering, and information extraction. Ninety-two people responded to the survey. As half of the questions were optional, not all questions were answered by all participants. Among the 92 respondents, 69 completed the survey (73.9%). Section 4 reports the results from the two stages.

<table>
<thead>
<tr>
<th>A. Resource Type</th>
<th>C. Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>A.1 Knowledge source</td>
<td>C.1 Subfield</td>
</tr>
<tr>
<td>A.1.1 Monolingual lexicon/ontology</td>
<td>C.1.1 Monolingual QA</td>
</tr>
<tr>
<td>A.1.2 Bilingual or multilingual lexicon/ontology</td>
<td>C.1.2 CLQA</td>
</tr>
<tr>
<td>A.1.3 Annotated corpus</td>
<td>C.1.3 CLIR</td>
</tr>
<tr>
<td>A.1.4 Un-annotated corpus</td>
<td>C.1.4 Automatic classification or clustering</td>
</tr>
<tr>
<td>A.1.5 Test collection</td>
<td>C.1.5 Summarization</td>
</tr>
<tr>
<td>A.1.6 Other</td>
<td>C.1.6 Information extraction</td>
</tr>
<tr>
<td>A.1.7 Unclear</td>
<td>C.1.7 Other</td>
</tr>
<tr>
<td>A.2 Software system</td>
<td></td>
</tr>
<tr>
<td>A.2.1 Information retrieval system</td>
<td>C.2 Proportion of use</td>
</tr>
<tr>
<td>A.2.2 Natural language processing tool</td>
<td>C.2.1 Entirely used</td>
</tr>
<tr>
<td>A.2.2.1 Information extraction tool</td>
<td>C.2.2 Partly used</td>
</tr>
<tr>
<td>A.2.2.2 Parts-of-speech tagger</td>
<td>C.2.3 Unclear</td>
</tr>
<tr>
<td>A.2.2.3 Parser</td>
<td></td>
</tr>
<tr>
<td>A.2.2.4 Morphological processor</td>
<td></td>
</tr>
<tr>
<td>A.2.2.5 Other</td>
<td></td>
</tr>
<tr>
<td>A.2.3 Machine translation tool</td>
<td></td>
</tr>
<tr>
<td>A.2.4 Classifier</td>
<td>C.3 Specific purpose of use</td>
</tr>
<tr>
<td>A.2.5 Other</td>
<td>C.3.1 Stop word removal</td>
</tr>
<tr>
<td>A.2.6 Unclear</td>
<td>C.3.2 Indexing</td>
</tr>
<tr>
<td>A.3 Other</td>
<td>C.3.3 Query expansion</td>
</tr>
<tr>
<td>A.4 Unclear</td>
<td>C.3.4 Document retrieval</td>
</tr>
<tr>
<td>B. Resource Acquisition</td>
<td>C.3.5 Disambiguation</td>
</tr>
<tr>
<td>B.1 Freely available</td>
<td>C.3.6 Morphological analysis</td>
</tr>
<tr>
<td>B.1.1 Web</td>
<td>C.3.7 Syntactic parsing</td>
</tr>
<tr>
<td>B.1.2 Other</td>
<td>C.3.8 POS tagging</td>
</tr>
<tr>
<td>B.2 Purchased</td>
<td>C.3.9 Term translation</td>
</tr>
<tr>
<td>B.3 Self-developed</td>
<td>C.3.10 Named entity identification</td>
</tr>
<tr>
<td>B.4 Other</td>
<td>C.3.11 Corpus/dictionary/ontology/database construction</td>
</tr>
<tr>
<td>B.5 Unclear</td>
<td>C.3.12 System training</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 1. Coding scheme for the use of resources in IIA

4. Results and Findings

The codes derived from the content analysis were examined using both frequencies of and relations between variables within the analysis. We also examined relations between certain categories, such as the frequency of use of resources and their availability, and the relations
between subfields and the number of resources used. Particular categories, such as category B.11, Freely Available Web Resources, were further analyzed. The survey results questions were analyzed using descriptive statistics to verify the findings from Stage One and to understand how the IIA research community acquires and shares resources.

4.1. Resources used by QA and CLIR Research

In Stage One, our analysis identified 290 resources from the 145 sample papers. Of the 290 resources, 118 (40.7%) belong to knowledge sources and 171 (59.0%) are various software systems. One resource is coded as A.3: “Other.”

Knowledge sources include various lexicons, dictionaries, thesauri, annotated or non-annotated corpora, and test collections, as depicted in Figure 2. The Labels in the Figure are the categories as shown in Figure 1. For example, A.1.1 is monolingual lexicon/ontology. Software systems include information retrieval systems, machine translation systems, and various natural language processing (NLP) tools. Their distribution is depicted in Figure 3. Figure 3 shows natural language processing tools are commonly used in IIA.

Due to the limited sample size of the analysis, the duplication rate of the use of the same resources in different papers is low. Almost three fourths of the resources (72.1%) were only mentioned in one single paper or poster. However, certain resources were more frequently used. WordNet (http://wordnet.princeton.edu/) was turned to be the resource mostly used for both QA and CLIR. In many papers, WordNet was only partially drawn upon, to build semantic taxonomy or to identify features such as parts-of-speech of words and phrases. Other resources that were used by more than 5 papers include: Google, LUCENE, TREC data collection, SysTran, BabelFish, ChaSen, Snowball, Lemur toolkit, Free Translation, EuroGOV collection, Porter stemmer, Minipar, TreeTagger, WorldLingo, Google Translator,
BBN’s IdentiFinderTM, IR-n system, and LDC (Linguistic Data Consortium) dictionaries. Among these resources, three stand as knowledge sources (i.e., WordNet, TREC test collection, and LDC dictionaries). Others are software systems including six NLP tools, four IR systems or search engines, and six machine translation systems.

4.2. Characteristics of IIA Resources

The coding results discovered some characteristics of the resources used in the sample paper. More than half of the resources (59.2%, 170 resources) can be clearly identified as freely available. They can be accessed without charge on the Web (called Web resources) or can be obtained for free upon request. Specifically, Fourteen out of the 19 most frequently used resources are Web resources publicly and freely available for research purposes.

Resource sharing in IIA is common and in international scope. Resources are developed in different organizations/research groups in several countries. Some were produced with funding from government agencies. Most researchers have harnessed resources developed by others doing research or development. As presented in Figure 4, among the 145 sample papers, only 13 (8.9%) did not use any resources developed by others. All the rest have used one or more existing resources. Nearly half of the sample papers (47.6%) used between two and four resources generated by other individuals or research teams.

![Figure 4. Frequency distribution of resources identified in the sample papers](image)

Research in subfields uses slightly different types of resources. For example, the resources most typically used in CLIR are machine translation tools and bilingual/multilingual lexicons, while those usually used for monolingual QA are information retrieval (IR) systems. Scholars consider translation the most challenging issue facing CLIR systems aiming to perform as optimally as their monolingual counterparts, which probably
explains why machine translation tools and bilingual/multilingual lexicons were the most frequently-used resources for CLIR. And IR is usually the first step towards automatic question answering.

IIA researchers are quite creative in making use of available resources. A resource can be used quite differently from its original intended use. Our coding process demonstrated that resources can be used in many different processes in the research and that, therefore, the 17 third-level categories under Category C.3, Specific purpose of use, were not sufficient to group and specify various kinds of resource usages. We eventually had to redevelop that category (C.3) and used the new categories to recode the sample papers. We came up with 12 second-level categories, and 29 third-level categories. The results reveal that text processing and translation are the two biggest uses of resources. Together they account for 39.9% of the total usages. Other important uses include system/algorithm testing (11.6%), information or knowledge extraction (9.9%), and question/query processing (7.1%).

4.3. Resource Acquisition and Sharing

The third research question “How do IIA researchers acquire and share resources?” is mainly answered through the analysis of the survey results in Stage Two.

We asked the participants to identify the channels in which they acquired the needed resources for their research and development. Sixty-seven out of seventy-two respondents (93.1%) employ Web search engines to locate resources for research and development purposes. Websites of related academic associations, major research centers, or resource consortia are the second most important sources. Also, researchers acquire resources from conferences or workshops they attended. Personal connections rank as the fourth most commonly used means of seeking appropriate resources. Based on the results, mailing lists are not preferred as an avenue for acquiring resources.

The next question is concerned with whether they can always find the right resources for their research, 38 out of the 49 respondents (77.6%) can sometimes find the right resources; 6 (12.2%) can always find them, and 4 (8.2%) can seldom find the right ones. One respondent reported that, on one hand, some resources exist but are not useful due to various obstacles; on the other hand, some resources do not exist yet. The researcher has to continue to search, or to develop them by him/herself. The obstacles include the quality of the resources and their compatibility with the researcher’s own research.

The survey also identifies important factors affecting researchers’ choices among similar resources. Money, reputation and ease of access rank as the top three considerations. Seventy-seven percent of the respondents sometimes can not afford the resources they would like. 81.6% of the respondents favor freely-available resources when they can find various
resources of the same type. Other factors include ease of installation, ease of learning, language of the programming, and quality of documentation.

5. Discussion

This section discusses issues related to the coding scheme, the definition of IIA and its subfields, and possible solutions to resource sharing problems.

The coding scheme is a useful outcome of this study, since it can be used beyond content analysis. It could help organize IIA resources if a resource management system were to be built. Resources organized by their use situation will help IIA researchers to quickly find the resources they need. However, our current coding scheme is not perfect even after the revision as reported in Section 4.2. For example, we had difficulty differentiating an unannotated corpus and a test collection, because the test collection contained a document collection which could be used separately in some circumstances as an unannotated corpus. Also, the revision to Category C.3 was based on the 145 sample papers in QA and CLIR. The category may need to be further expanded or modified when it applies to other fields.

The term Intelligent Information Access (IIA) has been used by professionals in Information Science, Computer Science, and Computational Linguistics, but has not been well defined. Especially unclear is what subfields IIA encompasses. Here we define two basic characteristics of IIA: (1) The purpose of IIA is to improve information access regardless of the format, media, and type of the information; and (2) IIA technologies should apply human knowledge or human-like understanding of information to achieve high performance. This is a quite expansive definition. The purpose is to ensure that the term could cover appropriate areas that relate to information access. The IIA community could be better defined if taxonomy of the subfields of IIA could be established.

Based on the survey results in Section 4.3 and our review of available resource-sharing channels in Section 2, we believe that the IIA community will benefit from an online resource management system. A resource management system should be developed based on a well-established knowledge-management model so that it facilitates knowledge creation and sharing. This idea reflects the expectation of IIA researchers. In our survey, we asked a question concerning a repository of IIA resources. 38 (79.6%) out of 48 respondents think that a repository of IIA resources would be useful, and 71.8% of those researchers expressed their willingness to share the resources they developed through such a repository.

6. Conclusion and Future Work

This exploratory study aims to understand resources and resource sharing among researchers and developers in Intelligent Information Access (IIA). The investigation discovered the
types of resources shared by the researchers for two important IIA subfields and achieved a
general understanding of the characteristics of the resources and resources shared in IIA.

Future research will focus on designing a resource management system (RMS) model for
IIA based on well-established knowledge-management models. We are very interested in the
KM Process Framework presented by Bukowitz and Williams (1999). The tactical part of the
framework defines the process of how an individual can interact with the KM system:
• Get – a knowledge worker uses the system to find information he/she needs. In the
  context of IIA, this will be the process of an IIA researcher searching the resource
  management system for needed resources;
• Use – the process of making use of existing knowledge to create new knowledge. In the
  context of IIA, this will be the process of IIA researchers conducting their research and
  development to create new IIA systems or resources, which in turn become resources in
  the system;
• Learn – this process involves learning from each other to promote knowledge and
  competitive advantage. In the context of IIA, this will be the process of IIA researchers
  interacting with each other in order to learn new resources and of new needs for resources;
• Contribute – this process gets the knowledge workers to contribute to the organizational
  knowledge bases. In the context of IIA, this will be the process of IIA researchers
  contributing to the resource management system to share their resources.

In general, the above tactical processes define the functions a resource management
system should have: the system should allow the IIA researchers to search, contribute and
discuss resources.

The strategic component of the framework provides guidance on steps of building and
managing the resource management system:
• Assess – this process identifies the needs for current knowledge management. In the
  context of IIA, this will be the user needs assessment and the strategic planning of the
  resource management system;
• Build and Sustain – this process includes the design and maintenance of the knowledge
  bases. In the context of IIA, this will be the actual design, development, and maintenance
  of the resource management system;
• Divest – this process evaluates the organizational knowledge base and makes appropriate
  adjustments. In the context of IIA, this will be the process of continuous evaluation of the
  content and the functions of the resource management system.

Even though the framework by Bukowitz and Williams was intended for organizational
settings, it could be well applied to the designing, developing, and maintaining of a resource
management system for IIA. Future research will further examine the application of this
framework to resource sharing and management in the context of IIA. We will focus on the
most challenging processes in the development and maintenance of the resource management
system and explore the solutions to these challenges in order to develop a successful resource management system for the IIA community.

References


