KNOWLEDGE MEDIA KMi INSTITUTE

Experiences of Two Task Driven User Studies of Hypermedia Information Systems

Tech Report kmi-05-4
May 2005

Victoria Uren, Philipp Cimiano, Simon Buckingham Shum and Enrico Motta

The Open University

Experiences of Two Task Driven User Studies of Hypermedia Information Systems

Victoria Uren[1], Philipp Cimiano[2], Simon Buckingham Shum[1], Enrico Motta[1]

Knowledge Media Institute, Open University, Milton Keynes, UK
(v.s.uren,s.buckingham shum, e.motta)@open.ac.uk

AIFB, University of Karlsruhe, Karlsruhe, Germany
cimiano@aifb.uni-karlsruhe.de

Abstract. We present two small scale user studies of hypermedia information systems: a hypermedia discourse system designed as an environment for researchers to summarize and share key ideas from research papers as a claim network, and a web browser plug-in which annotates terms related to a selected ontology on the fly. The first study investigated whether a claim network created by one user could help others learn about a domain. The second study investigated whether information extraction techniques for identifying extra domain terms enhanced the system. We discuss the strengths and weaknesses of these studies and the extent to which they achieved their goals.

1 Introduction

In our on-going research on innovative hypermedia and web-based systems, we often encounter the problem of having to evaluate the performance, or added value, of systems. Few of our systems are retrieval algorithms that could be evaluated using recall and performance metrics, even if suitable test-beds existed, and they are usually under rapid development, making large-scale user studies inappropriate. Recently, we have investigated specific questions about key areas of system performance through task driven user studies on a relatively small scale. By task driven we mean that the user's interaction with the system is directed by giving them a specific task to accomplish. These studies have allowed us to validate key assumptions about systems while continuing with active development. We present our experience with two of those studies in this paper.

2 Can People Understand Claim Networks?

The hypermedia discourse system investigated in this study has at its core a typed, directed network which we call a claim network. To build these networks, researchers identify the concepts in research articles which they find significant, instantiate them as nodes in a network with brief text summaries and link the concept nodes to other concepts in the same paper or elsewhere in the literature, using typed links. The link labels are based on a discourse ontology of rhetorical moves made by authors when building up arguments in papers, e.g. *is similar to*, *refutes* and *is capable of causing*. The construction of such networks is known to be helpful to individuals involved in making sense of their reading e.g. [1]. However, our ambition for the system goes beyond individual sense-making. Our vision is that such systems could provide a shared window onto distributed information resources, such as the papers in a digital library, with collaborating researchers contributing claims and counter claims to build a shared, searchable discourse space[2]. In this study we addressed the open question of whether one user could interpret a claim network constructed by another user.

2.1 Method

To determine whether the claim network could communicate information we compared the performance of two groups on a factual questionnaire based on topics which had been described in both a claim network and a brief written review. The participants in the study were six research students. None of them had prior, in-depth knowledge of the topics in the literature selected for the study. Half the group were engaged in research related to discourse mapping and literature analysis. These three were all familiar with the discourse ontology and the ideas underlying claim networks but were not particularly familiar with the tools for searching the networks (called ClaimFinder and ClaimMaker[2]). These students were assigned to the *Claim Network* group. The remaining three students were assigned to the *Written Review* group. It was not considered detrimental to the study to use the students with knowledge of the principles of claim networks, since we wished to investigate a scenario in which the basic ideas and instrumental operations were known (just as members of the Written Review group were familiar with reading, pens and paper).

A testing station was set up with the Camtasia screen capture tool (http://www.techsmith.com/products/studio/default.asp) to record the participants' interactions with the tools and their verbal comments. Participants were accompanied by an observer who could assist with any general queries and who also provided someone to "think aloud" to. The questionnaire was presented on screen. The Claim Network group was given a Microsoft Internet Explorer browser with links set up to both ClaiMaker and ClaimFinder. The Written Review group had an open Microsoft Word document containing the review, plus a hard copy version since many people prefer to read on paper.

With the setup described above we were able to gather a number of different kinds of data: how long it took to answer each question, any comments participants made (to get a qualitative view of their experience of the system), for the Claim Network group which of the search functions they used (giving a guide to which features of the tool were working well), and finally their answers to the questions.

2.2 Time to Answer Questions

We timed how long the two groups of participants took to complete the whole exercise and the proportion of their total time each person spent on each question. With such small test groups we knew that average timings were not going to be statistically significant but looking at proportional times gave us some insight into the relative difficulty of particular questions.

Actual time taken. Actual times (see Table 1) gave us a guide to the relative difficulties faced by the two groups in answering the questions. The Written Review group was clearly able to answer the questions faster than the Claim Network group. This was to be expected because the Written Review group was far more familiar with their task, essentially a reading comprehension test, than members of the Claim Network group. Also we noted that the Claim Network group gave more "thinking aloud" contributions, which was an additional task. The reluctance of the Written Review participants to think aloud may stem from the strong habit of reading silently. The great variability in the times taken by the Claim Network group seemed to be attributable to personal style rather than different levels of competence with the tools, in particular, the slowest participant had a very analytical approach to both the questions and the data in the claims.

Table 1. Total time spent on the exercise by each participant

Participant – task	Approx. time in mins
A – Network	54
B – Network	78
C – Network	183
Mean Network	105 (var 69)
D – Review	56
E – Review	36
F – Review	38
Mean Review	43 (var 9)

Proportion of Time per Question Differences in the proportion of time spent by the groups on each question indicate that one of the two media has an advantage answering that question. Figure 1 summarizes this data for the two groups. For most of the questions there was no indication that either group of participants found any question harder than the other. For Question 4 the Written Review group found it easier to answer the question than the Claim Network group, whereas for questions 7 and 9 the situation was reversed.

We do not have the space here to look at the questions in detail as we did for the original study. To summarize, we found that where there were differences between them they came from particular affordances of the two media. For example question 7 asked the users to find three properties of a certain thing. The network group had an advantage for this task because the information in it had already been decomposed into concepts. The Written Review group had to mentally apply a chunking procedure to the written text to obtain three properties.

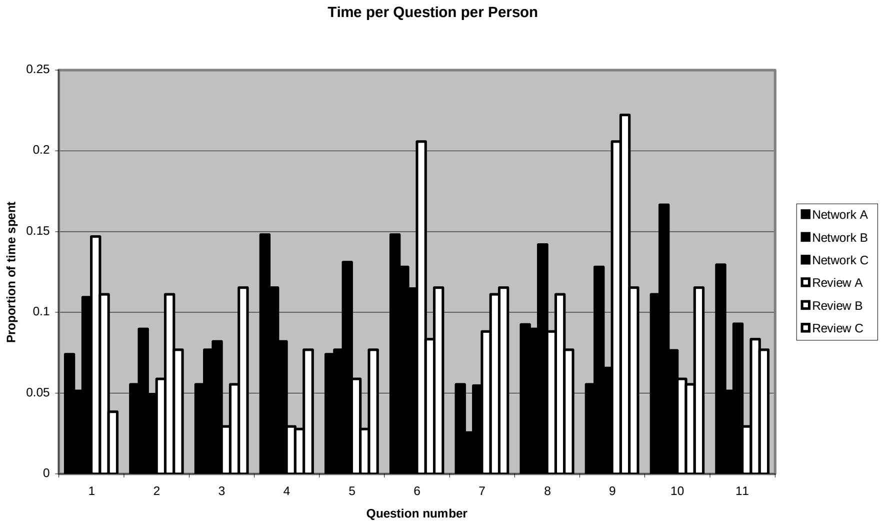

Figure 1. Proportion of time spent to answer questions compared for the Claim Network (black) and Written Review (white) groups

2.3 Tool Usage Patterns

We also used the recordings of the Claim Network group's sessions to assess how the functions of the search tools were used by counting the numbers of search actions of particular types. From this data we could see which search features had proved most useful. For example, a feature called the *Anchor* icon (42% of actions) was the most used. *Anchor* selects a concept to be the focus of a search and displays it with all the concepts that have claims linking in or out of it. Use of *Anchor* turned out to be part of a dominant searching strategy, which was to perform a simple keyword based search called *Find* (19%) to locate the topic required and then to explore the local region of the network. This data was invaluable in our ongoing system development.

2.4 Advantages and Disadvantages

A positive outcome of this study was that when we examined the participants' answers we found no difference in quality between the two groups. Therefore it demonstrated that the research students were able to understand this particular Claim Network using our search tools. They took longer to answer the questions than their colleagues with the written review, but we are not concerned by this as the skill levels of the two groups were different.

A weakness of the study was its dependence on two parallel artifacts in different media, the claim network and the written review. While we did our best to ensure the questions could be answered from either of these, it is not clear how one could ensure that two different media actually contain equivalent information. When we found differences between the difficulties encountered by the groups on particular questions they turned out to be features of the artifacts. For instance, if the three properties required to answer question 7 had been presented in the text as bullet points it is unlikely that the Claim Network group would have had an advantage.

We encountered some operational problems with our method. It depended very heavily on extracting data from Camtasia movies. This turned out to be a time consuming process. Each movie had to be watched several times to extract each kind of data and a considerable amount of "rewinding" was sometimes needed to be certain about what had happened. For example when determining exactly which action had been taken to produce a particular view of the claim network. Our timings also had to be approximate, e.g., we only timed questions to the nearest minute.

3 Does Information Extraction Enhance a Browser Plug-in?

In this second user study, we compared the performance of three groups of users on two fact finding tasks where the database they searched was the KMi Planet[1] news server. For two of the groups the basic server was supplemented with a semantic browser plug-in called Magpie [3]. Magpie allows users of web-based documents to interpret content from different conceptual perspectives by automatically generating annotations corresponding to a particular ontology as a semantic layer over the actual content of the document. Thus it can provide semantic web services for documents with no semantic mark up.

[1] http://sixfields.open.ac.uk/kmiplanet/

The end-user part of the Magpie framework is a browser plug-in. The user can choose an ontology and toggle categories of knowledge via buttons presented in a toolbar. Selecting a button highlights items in the text that are relevant to the chosen category. The user can access a menu with relevant web services for each item (this functionality was not used in the study).

These dynamic annotations are generated using a lexicon which relates each concept in the ontology to the various text strings by which it is commonly represented. The lexicons are currently produced by domain experts. We would like to automate this costly process and information extraction is an obvious method to test. Information extraction finds salient entities in texts. These might, for example, be the names of companies involved in merger negotiations, or the time at which an event took place. Our aim in this experiment was to see whether or not lexicons generated in part by information extraction improved the performance of Magpie users.

3.1 Method

The performance of three groups of participants, A (control), B (Magpie/AKT) and C (Magpie/AKT ++) were compared on two fact retrieval tasks. The database the participants searched was an online newspaper (Planet News) featuring events at the research institute where they worked. Planet News incorporates a Main News page, showing recent stories, News Archive pages, which have a reverse chronological listing of all the stories and a Search option which permits simple keywords searches.

The control system, used by Group A (control), was the Planet News interface with no additional features. Group B (Magpie/AKT), used the same interface augmented with the Magpie system using a pre-existing hand built lexicon based on the AKT reference ontology of academic life (http://www.aktors.org/publications/ontology/) with four upper level categories: Person, Project, Research-Area and Organization. These categories are instantiated in the Magpie system as 4 buttons which highlight entities of the selected type. Group C (Magpie/AKT++) also used Planet News augmented with Magpie but this time with a lexicon, called AKT++, built from three sources: the AKT ontology, entities extracted from the news stories by an named entity recognizer called ESpotter [4] and entities extracted from the news stories by an a web based entity recognizer and classifier called PANKOW [5]. This ontology had nine upper level categories: Person, Project, Research-Agenda, Organization, Place, Event, Politician, Technology and Company, so it was richer both in terms of content and in its organization.

The participants were a mix of research students and non-doctoral researchers. They all had web-searching skills and knowledge of the subject domain. Group A had six participants, Group B, seven, and Group C, seven. Each participant was given a demonstration of the interface and was then asked to do two timed fact retrieval tasks in succession, which they completed in the presence of an observer. In the “People” task they had to compile a list of important people who had visited their institute. In the “Technology” Task they had to compile a list of technologies, either in-house or external, used in their institute’s research projects. Their answers had to come from the Planet News stories and they were allowed 10 minutes to complete each task. The participants recorded their answers by cut and pasting items from the stories into a text file. Their interactions with the interface were recorded using Camtasia Studio.

We obtained the following results from this experiment: summary statistics from an analysis of the quantity and quality of items retrieved by each group, an analysis of how many of the items each group retrieved were in one of the two lexicons, and an analysis of interactions with the tools acquired from the Camtasia movies.

3.2 Retrieval Performance

We examined whether the Magpie annotations improved the number and quality of items the participants retrieved in the time available. For this we needed an independent assessment of each item that was given as an answer. Two cumulated lists were produced of the 134 people and 133 technologies that appeared in answers. These lists were presented to an impartial assessor, who was a long serving member of the institute and who had not been involved in the design or running of the experiments. He rated each item 0, 1 or 2. The total value of scores that he applied was 94 for People and 140 for Technologies.

Scores for each participant are the sum of the scores for all their answers. Mean scores for the three groups on both tasks are presented in Table 2. Both the groups using Magpie achieved higher scores for both tasks than the control group. Group B (Magpie/AKT) did best on the People Task, whereas Group C (Magpie/AKT++) did best on the technologies task. Further analysis is needed to explain the relatively low score for group C on the People task.

Table 2. Mean scores for the People and Technologies tasks

Task	Group A (Control)	Group B (Magpie/AKT)	Group C (Magpie/AKT++)
People	13.2	15.3	13.7
Technologies	19.2	23.4	26.7

3.4 Answer Coverage

In this part of the analysis we compared how “good” the two lexicons (AKT and AKT++) were for answering the questions. We determined how many of the items the participants copied into their answers were in one of the two lexicons. We included all three groups, whether they actually used the Magpie system or not and everything the participants pasted into their answers irrespective of whether the items were judged to be correct. Taking this approach gave us a bigger sample to work with.

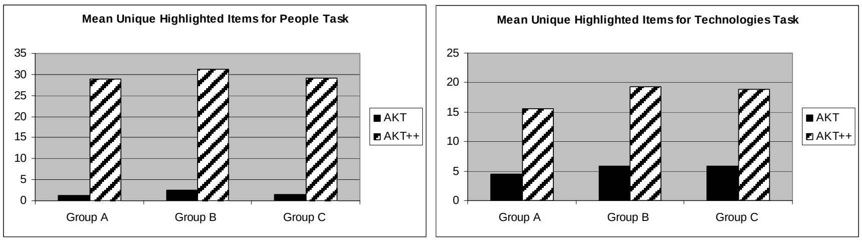

Figure 2. Answer coverage for the People and Technologies Tasks

For all three groups and for both tasks we found that the AKT++ lexicon highlighted more items per answer than the AKT lexicon alone (see figure 2). This was to be expected; the AKT++ lexicon contained the AKT lexicon so, unless ESpotter and PANKOW were extracting nonsense, we would expect AKT++ to contain more answers. However the differences are substantial (for all six cases they were significant at the 2.5% level in two-tailed T-tests). Therefore we conclude that the information extraction tools were fit for the purpose of populating lexicons for these two tasks.

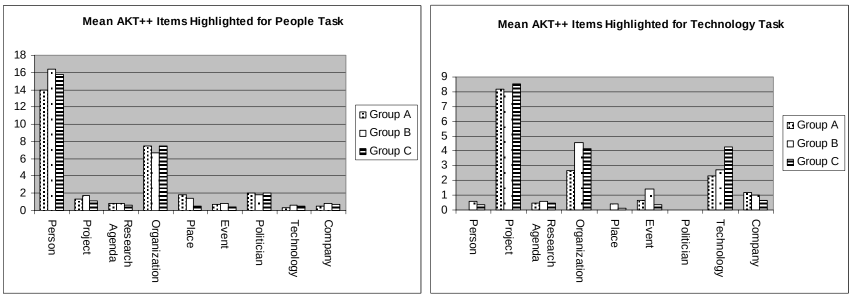

Figure 3. Breakdown of AKT++ highlighting by type for the People and Technology Tasks

A more interesting question was whether the extracted entities in the AKT++ lexicon were automatically assigned to appropriate types? Figure 3 presents the answers broken down by type for the AKT++ highlighting. For the People task, the answers were mainly recognized as either people or organizations (some participants included visitors' affiliations in their answers). For the Technology task (Fig. 4), the answers were split between Project (which identifies many of the institute's own technologies), Organization (were a technology contains the company's name) and of course Technology. Overall we were satisfied that the AKT++ lexicon was relating items to types which would help users attempting to answer these questions, i.e., the classification aspect of information extraction was also fit for purpose.

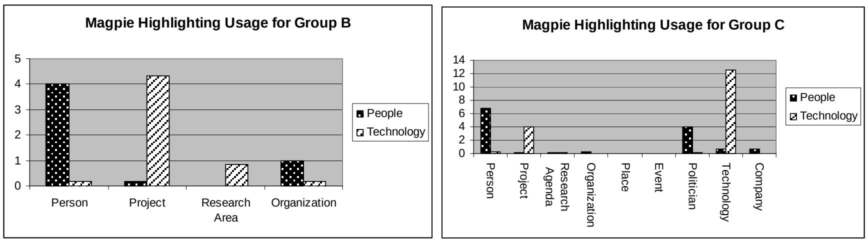

Figure 4. Magpie highlighting usage for the People (dark) and Technology (light) tasks for Group B (left) and Group C (right)

3.5 Movie Analysis

In Magpie the highlighting had to be refreshed for each new document that was viewed. Therefore we were able to judge how useful the participants found different highlighting options by seeing whether they used them repeatedly or whether they gave up on them after a few unfruitful trials. The Camtasia movies recorded during the experiment were analyzed to see how often the participants selected each of the Magpie highlighting options. Figure 4 presents the mean usage of the different highlighting options for Group B and Group C (Group A did not use Magpie). The most used highlighting options for Group B are Person and Project. For Group C the most used options are Person, Project, Politician and Technology.

3.6 Advantages and Disadvantages

As for the claim network study, this study achieved its main aim. We were able to show that the information extraction based AKT++ lexicon gave better results than the hand built AKT lexicon.

Once again hindsight suggests omissions in our method. In our analysis of the use of the AKT++ lexicon we made no distinction between the manual annotations, PANKOW annotations and ESpotter annotations. This did not concern us for this particular experiment as we were examining the hypothesis that NLP tools could add value in general and therefore built the most comprehensive lexicon available to us. However, if we had distinguished the sources of the annotations we could have also compared annotations produced with different NLP approaches.

Practical considerations led to the introduction of confounding factors in the experimental design. We compared the pre-existing lexicon AKT with the best lexicon we could make, AKT++, which contained both manually and automatically created entries. It could be argued that we would have got clearer results be comparing a purely manual lexicon and a purely extracted one. However we chose the realistic mixed lexicon, since in an operational environment nobody would throw away a good but incomplete knowledge base; they would enhance it.

Although we took a quantitative approach to measuring performance, we did not look at precision and recall since we did not have the manpower to create a gold standard. However we have noted since completing this experiment that our method of using an independent assessor to rate answers after collection parallels that used for some TREC tracks, e.g. [6]. We will investigate whether we can obtain recall and precision data by this route.

The retrieval performance and answer coverage methods gave useful results with moderate effort. For a larger study both could be automated by writing scripts to analyze the participants' answers. The Camtasia movies also gave useful information but, as for the previous experiment, analyzing them was time-consuming. Overall the cost/benefit ratio for analysis time and data obtained was more favorable for this experiment than the evaluation of the claim network.

An advantage of the comparatively simple nature of this task was that we had a bigger pool of potential participants than for the claim network study. The primary benefit of this was that we could calculate significance for some of the results. Controlling the cost/benefit of data collection was vital for doing a larger study; it would not have been feasible to base all our data collection on the Camtasia movies for 20 participants as we had when there were only 6. A secondary benefit arose from our decision to limit the time available in that it not only motivated the participants to concentrate on the task in hand but also brought out the competitive instincts in some of them. After the experiments some people were very interested to know who had achieved the highest scores in their group.

4 Lessons Learnt

We face the challenge of having to devise new kinds of studies to evaluate novel hypermedia systems. We have found that small scale user studies that match task to key questions are helpful. However design is crucial. Ideally tasks should be sought which can be completed by a statistically significant number of participants and which can be analyzed using quantitative performance measures with reasonable cost/benefit at the data analysis stage. The tasks we set for the second study were relatively simple fact finding exercises. However, as Semantic Web systems grow more complex we expect them to be able to perform more complex operations. Therefore we will need to develop user studies that can evaluate performance for more complex tasks.

Our experience of the Techsmith Camtasia Studio screen capture software has shown that it has low set up costs but high collection costs. While it undoubtedly has advantages as a means of preserving a session for qualitative analysis, we need to investigate methods, such as keystroke logging, for recording participants' actions more efficiently in order to study more complex tasks.

Acknowledgements

This work was funded by the Advanced Knowledge Technologies (Magpie/AKT) Interdisciplinary Research Collaboration (IRC), the Designing Information Extraction for KM (Dot.Kom) project. AKT is sponsored by the UK Engineering and Physical Sciences Research Council under grant number GR/N15764/01. Dot.Kom is sponsored by the European Commission as part of the Information Society Technologies (IST) programme under grant number IST-2001034038.

We also thank all the participants in the experiment, Arthur Stutt for providing the impartial assessment of the retrieved items, Jianhan Zhu for ESpotter support and Laura Goebes for her help with analyzing the movies

References

1. Novak, J.D., Gowin, D.B., *Learning How to Learn*. 1984, Cambridge: Cambridge University Press.
2. Simon J. Buckingham Shum, V.U., Gangmin Li, Bertrand Sereno and Clara Mancini, *Modelling Naturalistic Argumentation in Research Literatures: Representation and Interaction Design Issues.* International Journal of Intelligent Systems, Special Issue on Computational Modelling of Naturalistic Argumentation (Eds.) Chris Reed and Floriana Grasso, 2005: p. (to appear).
3. Domingue, J.B., Dzbor, M. *Magpie: Browsing and Navigating on the Semantic Web*. in *Conference on Intelligent User Interfaces*. 2004. Portugal.
4. Zhu J., U.V., and Motta E. *Adaptive Named Entity Recognition for Web Browsing*. in *Workshop on IT Tools for Knowledge Management Systems at WM2005 Conference*. 2005. Kaiserslautern, Germany.
5. Cimiano P., H.S., Staab S. *Towards the Self-Annotating Web*. in *13th International World Wide Web Conference, (WWW 2004), May 17-22, 2004, New York, NY*. 2004. New York, NY.
6. Oard, D.W.a.G.F.C. *The TREC-2002 Arabic/English CLIR Track*. in *Eleventh Text REtrieval Conference (TREC 2002)*. 2002.