Aalto University
School of Science
Degree Programme in Computer Science and Engineering

Zhe Xie

From Exploration to Sensemaking:
an Interactive Exploratory Search System

Master's Thesis
Espoo, March 13, 2015

Supervisor: Professor Samuel Kaski, Aalto University
Advisor: Docent Jaakko Peltonen, Aalto University

Aalto University
School of Science

Aalto University
School of Science
Degree Programme in Computer Science and Engineering

ABSTRACT OF
MASTER'S THESIS

Author: Zhe Xie

Title:
From Exploration to Sensemaking: an Interactive Exploratory Search System

Date:	March 13, 2015	**Pages:**	54
Major:	Computer and Information Science	**Code:**	T-61

Supervisor: Professor Samuel Kaski, Aalto University
Advisor: Docent Jaakko Peltonen, Aalto University

The amount of data available online is increasing rapidly nowadays and enormous search systems exist to satisfy people's information needs. Many search systems have been designed for well-formed and explicit information needs but few for exploratory purposes. By exploration, we mean the information need is ambiguous initially and evolves during the search process.

There are mainly two parts within the thesis. In the first part, we will develop an interactive exploratory search system for the arXiv database, an open e-print archive for scientific articles. The part is mostly based on an initial study, in which a search engine, Scinet, based on an intent estimation model is proposed. With this search engine, users can direct their exploration by giving feedbacks to the estimated search intents, which are represented by relevant keywords. Intents are visualized and arranged into a radial layout where the radius measures relevance and the angle measures similarity. Users can drag a keyword closer to the center to indicate higher relevance or click on a keyword to assign full relevance and then the retrieved documents will be updated accordingly. Compared to the initial search system, a mind-map functionality is also added as a new feature. With this mind-map, users can temporarily store the keywords or titles that they find interesting. To verify the interactive exploratory ability, we have designed and conducted a small-scale experiment based on the arXiv dataset. Particularly, the keywords for arXiv articles are extracted by an automatic keyword extraction algorithm since most of the arXiv articles do not provide keywords by the authors.

For the second part, we investigate a potential novel functionality of the Scinet search system on a large database of scientific articles. The ability of the system to support information seeking was shown in previous publication. Here we propose that this system will also support sensemaking, namely, help users to make sense of the results. We suggest that this advantage arises because in Scinet, not only are intents estimated but also the relationships between them are indicated on the interface. In order to better support sensemaking, the new functionality is also added to the prototype system in the initial study. We believe that this search system will help people to better understand and interpret the search results.

Keywords: exploratory search system, interaction, visualization, sensemaking

Language: English

Acknowledgements

First of all, I would give me deepest appreciation to my advisor, Docent Jaakko Peltonen, who has funded me from his academy research fellow research funds[1]. I must appreciate the trust and admit that it was such an impressive experience to work with him. Jaakko has showed his great care and sense of responsibility to his student during the time when I was under his guidance. Moreover, I have to say that I have learned more than knowledge and skills from him but also the attitude to scientific researches.

Secondly, I would thank my supervisor, Professor Samuel Kaski, who has given the approval and support to the thesis topic. I would also thank Tuukka Ruotsalo and Antti Kangasrääsiö. Tuukka Ruotsalo has provided a strong support to us by explaining the architecture of the original system and providing suggestions to the design of the sensemaking user experiment. Specially, many thanks would be given to Antti Kangasrääsiö, without whom I could not proceed smoothly at the very beginning. His generous sharing of his work has greatly accelerated my working process.

Thirdly, I would like to express my gratitude to the Department of Computer Science and the Helsinki Institution for Information Technology (HIIT), who has provided the basic facilities to me. Moreover, I would thank the Re:Know project for the collaboration and the use of resources, such as servers, codes and data.

Finally, I would give my appreciation to my colleagues in the office room A319 who are so friendly and lovely. In the free time, we would have coffee and lunch together; we would also share our personal life and experiences. It was a wonderful time.

Zhe Xie

[1] Academy of Finland decision number 283311. The full project name is "Advanced machine learning methods for nonlinear dimensionality reduction, visualization, and exploratory data analysis based on multiple data sources"

Abbreviations and Acronyms

AKE	Automatic Keyword Extraction
TF-IDF	Term Frequency - Inverse Document Frequency
PCA	Principal Component Analysis
IESS	Interactive exploratory search system
POS tagging	Part Of Speech tagging
API	Application Program Interface

Contents

Chapter 1

Introduction

Information is of vital importance in the modern society and various information search systems are existing nowadays to fulfil people's information need. Even though the performance of these search systems, such as Google[1] and Baidu[2], are satisfying enough for common usage, the scientific community and the IT industry are still pursuing more accurate search results in order to provide better services. Unfortunately, information needs are complicated in essence and one important linchpin to success is to understand the users' intents or real needs more accurately. To simplify the problem, we can classify searches into two categories, direct search and exploratory search. Understanding these two different search behaviours will help us to better understand the users' intents.

For example, if someone is planning to go hiking near her/his home and not sure whether the weather will be appropriate, s/he can search the current weather information on the Internet with just one query. This is a direct search. However, information need would be much more complicated in other circumstances.

Imagine that a Chinese person is planning to visit Europe for several days. He knows little about Europe before and this time he thinks he must be well prepared in order to have a great time. Issuing a query like "how to have a good time in European countries" into a search engine, most probably will not provide him any satisfactory answers. Instead, he realizes that he needs to search for some more specific information, such as the flight information, what is most appealing season for tourists, the local securities and local transportations. Further, he will realize that some local history knowledge is also beneficial for better understanding the local culture, such as architectures and arts. This amount of found information could be over-

[1] www.google.com

[2] www.baidu.com, the most popular search engine in China

whelming if following the leads one after another. This is a typical feature of exploratory search. More formally, exploratory searching behaviours can be characterized by difficulty in expressing the search intent and the evolving information needs. Obviously, exploratory search is more complicated than a direct search. In this sense, special tools designed for assisting the exploratory search process are needed.

In order to support the exploratory search in a better way, the essence and characteristic of exploratory search must be studied. In the paper [12], the author defined a hierarchy of information needs, which corresponds to three search activity categories, including looking-up, learning and investigating. We deem that the exploratory search process is in essence a form of learning. This point of view is also supported in Freund et al. [2]. In the above examples, searching for weather information can be regarded as looking-up whereas searching for preparing an European trip just corresponds to the process of learning. Nevertheless, learning itself is still a broad concept and can be interpreted from different angles. According to the paper [2], learning can be characterized as knowledge acquisition, sense-making, interpreting and synthesizing. In this thesis, we typically will focus on the aspect of sense-making.

Generally, this master thesis has two main parts which have two different but related topics.

In Part I, an interactive exploratory search system (IESS) will be developed for a free scientific article database, arXiv. The IESS can be considered a extension of a former system based on a user intent estimation model. Utilizing this search system, the user's intents are estimated and visualized on an Intent Radar, a visual interface where the intents are organized according to their similarities between each other and the closeness to the user's estimated intents. A small-scale user experiment is performed as well. The experiment is aimed to verify the original conclusion that the user intent model or the interactive exploratory search system will significantly improve the user performance when conducting exploratory search tasks.

In Part II, we make the hypothesis that this IESS might help users better make sense of the search results in the exploratory process. We make this claim based on the observation that not only the estimated intents are visualized on the Intent Radar but also the relationships between the intents are indicated implicitly: the angles and locations between the intents imply their similarities. The meanings of exploration, learning and sensemaking in the thesis are interchangeable since exploratory search can be viewed as a process of learning and learning can be interpreted from the perspective of sensemaking. A user experiment will be conducted to test the sensemaking ability of the interactive exploratory search system, in which another variant system is created and set as a comparing system.

The thesis is organized as follows:

Chapter 2 will introduce and summarize the background work of the thesis. The contents include the interactive intent modelling, which is the theoretical foundation for the interactive exploratory search system in Chapter 4 and its performance based on the user experiment results.

Chapter 3 will present a literature review on the AKE algorithm and try to the find a state-of-art AKE algorithm. The IESS developed in Chapter 4 is built on the extracted keyword list of each article. Roughly speaking, the system performance depends on the quality of the extracted keywords. So the performance of the proper AKE algorithm would therefore play an important role in the overall quality of the search system.

Chapter 4 will introduce the development of the IESS for arXiv in detail. We will first give a introduction to the scientific database arXiv and then a detailed explanation to the contribution of the IESS, including the user interface and the data processing steps.

Chapter 5 will describe a small-scale user experiment that was conducted to verify the ability of the arXiv IESS to support exploratory search.

Chapter 6 will examine the sensemaking ability of the interactive exploratory search system. As stated previously, the IESS is initially designed for assisting the user to enlarge search scope and retrieve more relevant articles via the estimated intents shown on the Intent Radar. However, not only the estimated intents but also the relationships between the intents are shown on the Intent Radar as implied in the layout of the intents. To some extent, these relationships might further help the user to make sense of the information they see in search results and thus better understand the topics they are searching for. To verify the sensemaking functionality, a user experiment is designed and conducted, in which participants are required to undertake topic comprehension and comparison tasks.

In Chapter 7, we will conclude the thesis and highlight directions for the future work.

Chapter 2

Previous work

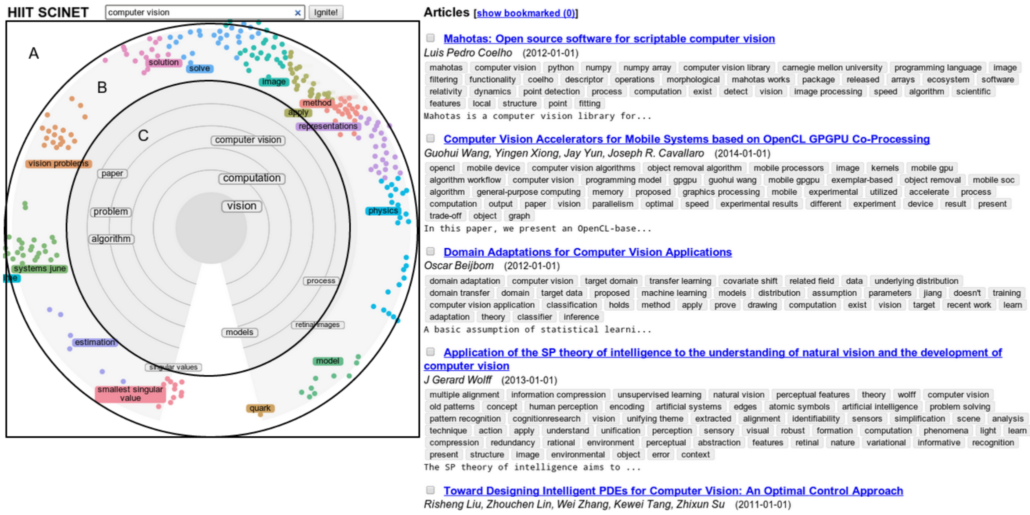
In this chapter, we will introduce the concept of interactive intent modelling and the results of the user experiment from the previous work. This interactive intent modelling algorithm provides the theoretical foundation for the previous interactive exploratory search system (IESS) called Scinet and for the work in this thesis. The user experiment results have showed that the IESS has a strong effectiveness of supporting users' exploratory search.

2.1 Interactive Intent Modelling

Most of the existing search techniques are designed for information retrieval tasks and evaluated by the quality of the retrieved documents. They assume that the users have a clear information need and have the ability to define their need very well. However, searching can be considered an exploration process as well, in which users want to learn or know new topics which they are no familiar with yet. In this case, the information retrieval task has become the challenge of how to assist users to search within an information space when they do not have a well-defined need yet. Ideally, the search system should be intelligent enough to understand users' intents and help them explore the information space.

The interactive intent model proposed in the paper by Ruotsalo et al. [17] is aimed for the information exploration challenge. In this model, users' search intents are estimated by the exploratory search engine and users' exploration will be guided by their feedbacks to the estimated intents. In the following, an example will be used to illustrate the system interface.

Suppose a user wants to learn about "computer vision" and find articles related to it. After the query, "computer vision", is issued to the search system, the system returns its response as shown in Figure 2.1.



Articles [show bookmarked (0)]

□ Mahotas: Open source software for scriptable computer vision
Luis Pedro Coelho (2012-01-01)
mahotas computer vision python numpy numpy array computer vision library carnegie mellon university programming language image filtering functionality coelho descriptor operations morphological mahotas works package released arrays ecosystem software relativity dynamics point detection process computation exist detect vision image processing speed algorithm scientific features local structure point fitting

```
Mahotas is a computer vision library for...
```

□ Computer Vision Accelerators for Mobile Systems based on OpenCL GPGPU Co-Processing
Guohui Wang, Yingen Xiong, Jay Yun, Joseph R. Cavallaro (2014-01-01)
opencl mobile device computer vision algorithms object removal algorithm mobile processors image kernels mobile gpu algorithm workflow computer vision programming model gpgpu guohui wang mobile gpgpu exemplar-based object removal mobile soc algorithm general-purpose computing memory proposed graphics processing mobile experimental utilized accelerate process computation output paper vision parallelism optimal speed experimental results different experiment device result present trade-off object graph

```
In this paper, we present an OpenCL-base...
```

□ Domain Adaptations for Computer Vision Applications
Oscar Beijbom (2012-01-01)
domain adaptation computer vision target domain transfer learning covariate shift related field data underlying distribution domain transfer domain target data proposed machine learning models distribution assumption parameters jiang doesn't training computer vision application classification holds method apply prove drawing computation exist vision target recent work learn adaptation theory classifier inference

```
A basic assumption of statistical learni...
```

□ Application of the SP theory of intelligence to the understanding of natural vision and the development of computer vision
J Gerard Wolff (2013-01-01)
multiple alignment information compression unsupervised learning natural vision perceptual features theory wolff computer vision old patterns concept human perception encoding artificial systems edges atomic symbols artificial intelligence problem solving pattern recognition cognitionresearch vision unifying theme extracted alignment identifiability sensors simplification scene analysis technique action apply understand unification perception sensory visual robust formation computation phenomena light learn compression redundancy rational environment perceptual abstraction features retinal nature variational informative recognition present structure image environmental object error context

```
The SP theory of intelligence aims to ...
```

□ Toward Designing Intelligent PDEs for Computer Vision: An Optimal Control Approach
Risheng Liu, Zhouchen Lin, Wei Zhang, Kewei Tang, Zhixun Su (2011-01-01)

Figure 2.1: The user interface for the IESS. Search intents are estimated and exhibited in a radar layout, Area A. This radar layout can be further divided into two circles, the outer circle, Area B and the inner circle, Area C. This picture is taken from the paper by Ruotsalo et al. [17].

On the left side is the Intent Radar, where the user is presented by a dark grey circle and predicted intents are visualized within the inner grey circles, Area C. The distance of an estimated intent to the center indicates its relevance to the current estimated search intent. The angles of different keywords indicate the similarity between the keywords. The future intent projections are also predicted and visualized within the outer grey circles, Area B. The future intents present potential directions the user may want to investigate in the future. The user can drag the keywords closer or further to the center to indicate his interest to estimated intents and the article list will be updated correspondingly according to the feedback.

On the right side is the articles list where user can mark their interested articles, read the abstracts and follow the link to find the full-content paper. For each article, its corresponding keywords are listed below the title.

As for the underlying algorithms of the interactive intent modelling, it consists of three parts, including: document retrieval, learning of the search intent and the optimization of the intent layout.

For the task of document retrieval, the search intent model will produce a keyword weight vector $\hat{v}$ which represents the estimated intents and then each document in the document set C will be ranked by the probability $\hat{P}(\hat{v}|M_{d_j})$ that the weight vector $\hat{v}$, which is regarded as an estimation of a ideal document, can be sampled from the document d_j of the document set.

The probability $\hat{P}(\hat{v}|M_{d_j})$ is given by:

$$\hat{P}(\hat{v}|M_{d_j}) = \prod_{i=1}^{|\hat{v}|} \hat{P}_{mle}(k_i|M_{d_j})^{\hat{v_i}}$$

and

$$\hat{P}_{mle}(k_i|M_{d_j}) = \frac{c(k_i|d_j) + \mu p(k_i|C)}{\sum_k c(k|d_j) + \mu}$$

where the $c(k|d_j)$ is the count of keyword k in document d_j, $p(k_i|C)$ is the occurrence probability of the keyword k_i in the collection C and μ is set to 2000.

After all the documents are ranked by the probability $\alpha_j = \hat{P}_{mle}(k_i|M_{d_j})$, a sample of the ranked documents will be presented to the users by Dirichlet Sampling instead of directly selecting the top ranked documents.

As for the task of learning the search intent, two types of search intents will be estimated: the current search intent and the alternative future intents. The estimated current search intent is in the form of a relevance vector $\hat{r}^{current}$ over a keyword set. In the interaction between the system and the user, the user will assign a score $r_i \in [0, 1]$ to a subset of the keywords, where $r_i = 1$ means the keyword is fully relevant to the user and $r_i = 0$ means the user has no interest in the keyword. Particularly, the keyword k_i will be represented as a $n \times 1$ vector in which each entry is the TF-IDF score calculated over the n documents. The relevance score r_i of a keyword k_i, $\hat{r}_i$, will be estimated as $k_i^T w$, where w is estimated from the LinRel algorithm [3]. Then the vector of keywords is selected according to the largest upper confidence bound for the relevance score $\hat{r}_i$ as the current search intent. The largest upper confidence bound is calculated as $\hat{r}_i + \alpha\sigma_i$, where α is the adjustment of the confidence level and σ_i is the upper bound of the standard deviation of $\hat{r}_i$.

The alternative future intents are represented as relevance vectors $\hat{r}^{future,l}$ where $l = 1, ..., L$. For each l, a pseudo-relevance feedback of 1 will be assigned to the lth currently shown keyword. Then a pseudo-feedback for a future potential search will be created as the combination of the pseudo-relevance feedback 1 to the lth keyword and the previous feedback from last iteration. After that, the lth future intent will be calculated similar to the current search intent.

After selecting the future intents, they will be laid out on the outer circle of Intent Radar Interface, which is determined by a probabilistic non-linear dimensionality reduction algorithm.

As mentioned above, each feedback l will produce an $\hat{r}_i^{future,l}$ for keyword k_i. As a result, keyword k_i will have L values of $\hat{r}_i^{future}$ in total, which forms a relevance score vector for it. Suppose that $\tilde{r}_i$ denotes the estimated relevance score vector of keyword k_i. Then $||\tilde{r}_i||$, the norm of $\tilde{r}_i$, will be used as the radius on the radar for k_i and $\bar{r}_i = \tilde{r}_i/||\tilde{r}_i||$ will be used as the direction of the keyword k_i on the outer circle. Intuitively, future intents with similar direction $\bar{r}_i$ should have similar angles α_i on the layout.

The similarity between keywords k_i and k_j is described with the concept of neighbours based on the direction. The neighbouring relationship between keywords, k_i and all its neighbours k_j then can be described by a neighbour distribution $p_i = \{p(j|i)\}$:

$$p(j|i) = exp(-||\bar{r}_i - \bar{r}_j||^2/\sigma_i^2) \cdot (\sum_{j'} exp(-||\bar{r}_i - \bar{r}_{j'}||^2/\sigma_i^2))^{-1}$$

where σ_i is set as in [20].

The neighbouring relationship for keyword k_i and all its neighbours k_j on the radar layout is described with distribution $q_i = \{q(j|i)\}$

$$q(j|i) = exp(-|a_i - a_j|^2/\sigma_i^2) \cdot (\sum_{j'} exp(-||a_i - a_{j'}||^2/\sigma_i^2))^{-1}$$

where a_i and a_j are the angles of keywords k_i and k_j on the layout.

Then optimal angle α_i for k_i is the one which minimize the divergence between the two distributions q_i and p_i. The divergence is measured by the total Kullback-Leibler divergence D_{KL} between neighbourhood for the angles and the neighbourhood for the directions, as $(\sum_s D_{KL}(p_i, q_i) + \sum_s D_{KL}(q_i, p_j))/2$. This divergence measurement is a function of α_i and optimal α_i will be obtained by gradient descent which minimizes the total divergence.

The layout of the estimated intents in the inner circle is determined by the corresponding intents in the outer circle because the future search intents are estimated based on the interaction to the current search intent. Particularly, for a current estimated keyword k_l, the radius of the intent in the inner circle is set to the estimated relevance $\hat{r}_l$ and the angle a_l is set to "the highest weighted mode of angles a_i of future keywords k_i, where the angle of each future keyword is weighted by the predicted future relevance $\hat{r}_i^{future,l}$".[17]

2.2 User Experiment and Results

In the paper by Ruotsalo et al. [17], a task-based user experiment was conducted to test the effect of supporting exploration. Over 50 million scientific documents were utilised to construct the dataset for the IESS. 30 graduate students with a background in computer science or a related field were recruited to be the participants and two post-doctoral researchers were recruited to design the tasks. During the experiment, the participants will try to search useful scientific documents to answer the questions in these tasks. Particularly, the tasks were defined as scientific writing scenarios in two scientific fields selected by the recruited experts.

Three aspects of the IESS are investigated, including: the user task performance, the quality of the displayed information and the interaction support for directing exploration. The user task performance is measured by the average score of the participants' written answers rated by the experts. Particularly, the documents are rated by the experts on three levels, including relevance, obviousness and novelty. The keywords are also rated on three levels, including relevance, general and specific. Then the quality of the displayed information is measured by precision, recall and F-measure calculated on the displayed documents and the keywords with respect to the expert ratings. The interactive support for directing exploration is measured both by the number and type of interactions and the percentage of different types of information displayed after different types of interactions.

Another two search systems are created as baseline systems to be compared to the Scinet IESS. One variant system is similar to the Scinet IESS. The only difference is that the estimated intents are not visualized on the radar layout but in a list format. The other variant system is a conventional typed-query based system, where no estimated intents are provided. For one participant, one system will be used for one single task in order to avoid learning effects. The results are summarized in the figure 2.2.

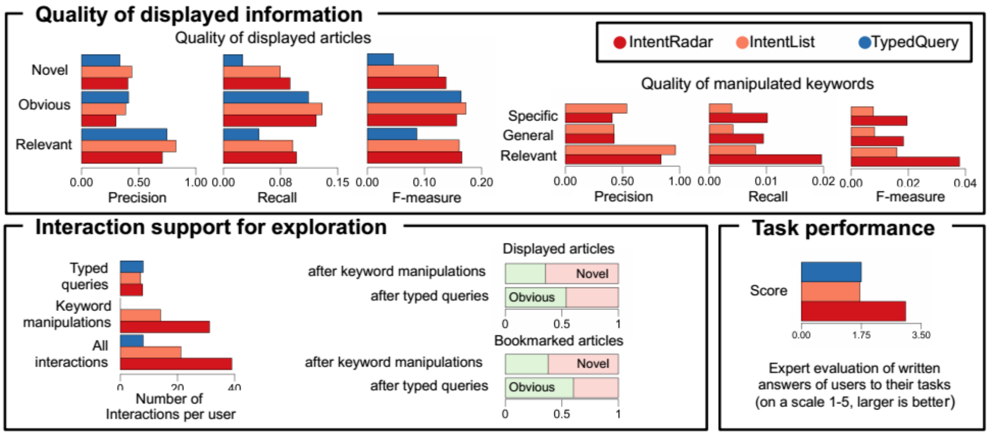


Figure 2.2: experiment results taken from the paper by Ruotsalo et al. [17].

All the three evaluation aspects are shown in the figure 2.2. From the figure, we can observe that participants using the system with the Intent Radar layout achieved better task performance; the quality of the displayed information from the IESS is significantly better than the quality from conventional search system. Moreover, it is also shown that the Scinet IESS would attract the user to interact with the exploratory search actively and after the interaction, novel documents are more easily exposed to the users.

Chapter 3

Keyword Extraction

In this chapter, we will explain the role and importance of the automatic keyword extraction (AKE) algorithm for our work and identify our practical expectations for the ideal AKE algorithm. We will present a literature review on AKE algorithms and then try to find the most suitable state-of-art AKE algorithm.

3.1 Motivation for AKE Algorithms

The Intent Radar of the IESS system and its corresponding probabilistic estimation algorithms as introduced in Chapter 2 are the realization of the interactive intent model proposed in the paper by Ruotsalo et al. [17]. In the interactive intent model, the user intents are represented by a keyword vector. The keywords themselves are assumed to be provided beforehand. It is clear that the quality of the keyword will determine the quality of the estimated intents and thus it plays an extremely important role for the whole search system. In the former work [17], the keywords used in the model is directly derived from the author-assigned keywords in each article. Unfortunately, most of the articles in the arXiv database that will be introduced in Chapter 4 do not provide author-assigned keywords in a direct way and we need to generate the keywords for each article by ourselves. The most feasible way is to extract keywords from the full text of the articles. This task is called automated keyword extraction and it is a research topic in the field of natural language processing. Fortunately, many algorithms have already been proposed and we should avoid reinventing the wheel. In the following, we will first state our expectations for the AKE algorithm and then conduct an literature review in order to uncover the most suitable state-of-art AKE algorithm.

Generally, we expect the ideal AKE algorithm should have the following features:

- The algorithm should have a competitive performance in extracting keywords.
- The algorithm should be relatively simple (e.g. fewer required features) so that the processing time on a large data collection will be tolerable.
- The algorithm should have a ready-made implementation to prevent coding from scratch.
- The algorithm should be unsupervised since author-generated or expert-generated keywords are usually unavailable for arXiv papers.
- The algorithm should need no domain-specific features since domain-specific features are mainly associated with a domain-specific glossary or keyphrase list. A glossary or keyword list is unavailable for the arXiv papers.

3.2 Literature Review on AKE Algorithms

The competition for outstanding automatic keyword extraction (AKE) algorithms at the Semantic Evaluation 2010 workshop [9] provides a precious opportunity to investigate the state-of-the-art algorithms for keyphrase extraction because many algorithms were competed against each other and all algorithms were evaluated under the same conditions. Nineteen AKE algorithms were submitted, among which thirteen algorithms belongs to supervised algorithms and six algorithms are unsupervised. The results showed that supervised algorithms outperform the unsupervised on average. However, KP-Miner proposed by EI-Beltagy et al. [6] which was ranked first within these 6 unsupervised algorithms had a very close performance compared with best supervised methods.

There are three steps in the KP-Miner algorithm, including candidate keyword selection, candidate keyword weight calculation and keyword refinement. In the candidate selection, two unique filtering conditions are applied: the first one is called the least allowable seen frequency, which will eliminate the phrases whose frequency is under this threshold, the second is the cut-off value which is the number of words after which if a phrase appears for the first time it will be filtered. In the candidate keyword weight calculation, a variate TF-IDF measurement is applied to which two boosting factors are added. In the keyword refinement, users are allowed to specify a number N of keywords that are expected to be returned.

Basically, the position information and a variate TF-IDF measurement are mainly used to weight the candidate keywords. Thus the system is very efficient from the point of view of computation. In the paper, the author also shows that the KP-Miner system outperforms another two widely-used keyphrase extraction systems, Extractor [19] and KEA [25]. Its better performance is also confirmed by You et al. [26].

Besides the algorithm of KP-Miner, new AKE algorithms are also proposed such as the algorithms proposed in You et al. [26], Sarkar [18] and Kang [8]. These methods are particularly interesting because they are all directly compared with KP-Miner and outperform it in the evaluations.

In the work by You et al. [26], an semi-supervised AKE algorithm is proposed. In the selection of keywords to be extracted, position features, (e.g. first occurrence), statistical features(e.g. phrase frequency) and granularity-related features(e.g. inverse document frequency difference) are taken into consideration. This is a supervised method because there are two parameters that need to be tuned by a training data set. These two parameters control the weight for their corresponding features. However, compared with other supervised method (e.g. [4],[11]), if we make a heuristic assumption about these two parameters, this approach also can be seen an unsupervised method and then one promising option for our purpose.

In the paper by Sarkar [18], a hybrid approach is proposed to extract keywords from medical documents. The proposed approach consist of two scoring strategies: one is based the phrase frequency and inverse document frequency and the other is based on domain knowledge for which a keyphrase list of 1940 keyphrases is created from the existing author assigned keyphrases collected from medical journal articles. Although this method outperforms the KP-Miner in the experiment,this method is not feasible for our situation since it requires a domain-specific knowledge. The reason for rejecting the method proposed by Kang [8] is similar to the one proposed by Sarkar [18] since a domain-specific glossary list is needed.

Besides the KP-Miner and the AKE system proposed by You et al. [26], there are other unsupervised systems as well.

In the work by Romero et al. [15], a thesaurus-based algorithm for support documents is introduced. Support documents refer to brief documents that help non-experts users learn the main concepts of any topics. They are usually domain-specific and only contain specific information rather than general or comprehensive information. Particularly, FAQ lists are mainly focused in this paper. Terms from Wikipedia are used to decide the score for each keyword candidate. Compared with other extraction system, including TF-IDF, Yahoo! Term Extractor, Wikify!, TextRank, and Longest common Substring, the proposed system performed much better in terms of F-scores.

In the work of Vidal et al. [21], keyword extraction is applied to web page content and used to provide more precise advertising service. Although the target of this algorithm is web content, it can still be generalized for general keyword extraction. Three Wikipedia-based keyword extraction methods are proposed in this paper. The first one can be regarded as a variation of the classic TF-IDF method, named Wiki-TF-IDF,in which the term frequency is the same, but instead of calculating the IDF from the document collection, the IDF is calculated from all the Wikipedia documents. The other two algorithms are actually the extension of the first one. After selecting the top N keywords from the previous algorithm, the categories of these keywords are identified and then all the terms belonging to the same categories are listed into a single file. After this, the Wiki-TF-IDF algorithm is applied to select the final keywords.

In the work of Wartena et al. [22], a new feature for keyword extraction is introduced, which takes correlations between words into consideration. The semantic relations between words are formalized with the co-occurrence distribution which is weighted average of the word distributions of all documents in which the word occurs. Then the co-occurrence distribution of a word can be compared with the document and the corpus distribution. In this way, the importance of the word can be evaluated. Typically, this feature is compared with the TF-IDF measurement and proved to outperform it.

In the work of Bohne et al. [5], the authors utilize a combination of heuristics to extract keywords from a single document. They regard the keyword extraction task as a ranking problem, in which words are ranked according to weights calculated from heuristics. These heuristics include TF-IDF, the Bernoulli model of randomness, the $\Gamma - Metric$ and the Laplace law of succession. Specially, the authors only consider the combination of two algorithms at a time so that the properties of each algorithm or heuristics are emphasized. In the combination of the weighting heuristics, the principal component analysis (PCA) is utilized. The eigenvalues calculated from the PCA are used to create a weighted combination of the heuristics weights. Particularly, the PCA is performed on the data table, in which the rows represent the candidate keywords and the columns represent the weights calculated from the heuristics.

In the work of Rose et al. [16], a rapid automatic keyword extraction algorithm is introduced. In the step of keyword candidate generation, the document text is splitted by stop words or phrase delimiters. Words between the split locations are considered a keyphrase as a whole. In the step of scoring the candidate keyphrases, the degree and frequency of the word within keyphrases and the ratio between them are considered. Degree of the word is defined as the frequency of the word when the word appears in a candidate keyphrase whose length is larger than 1. Then the score for each candidate keyphrase is computed as the sum of its member word score, for example, the ratio score. This paper also proposes a method to construct a stopword list. It proves that only considering raw frequency would cause content-bearing keywords to be included into the stopword list and a higher quality of stopword list could contribute the keyword extraction performance.

Although there exist other unsupervised algorithms for AKE, some of them are dedicated for certain field. For example, the method proposed by Romero [15] is aimed for support documents, such as tutorials and FAQ lists, which support the users to learn new concepts or skills and the paper by Vidal et al. [21] is aimed for web pages. More importantly, these methods can not be directly comparable since the evaluations provided by the authors are not performed on the same data set and the same rules.

3.3 Selection of a Keyword Extraction Algorithm for arXiv Data

Based on the research and the literature review, we prefer the KP-Miner as our first choice because it only needs a few features and its performance is rather competitive. Moreover, we can also utilize its existing API. However, we still may consider the algorithm proposed by You et al. [26] in the future due to its better performance.

Chapter 4

A Novel Search System for arXiv

ArXiv is an online scientific article database. In this chapter, an interactive exploratory search system (IESS) is developed for arXiv. Although there exist several search engines for arXiv, none of them support the interactive exploratory purpose very well. In the following, an overview of arXiv will be given and the IESS will be described in detail, including the user interface and the data processing steps.

4.1 Overview of arXiv

ArXiv is a highly-automated electronic archive and distribution server for research articles. Users can retrieve articles from arXiv and registered authors can submit and update their articles via a web interface. ArXiv was launched in August 1991. Since then, the monthly submission rates have been steadily increasing, which implies its growing popularity and importance. This growing trend is captured in Figure 4.1.

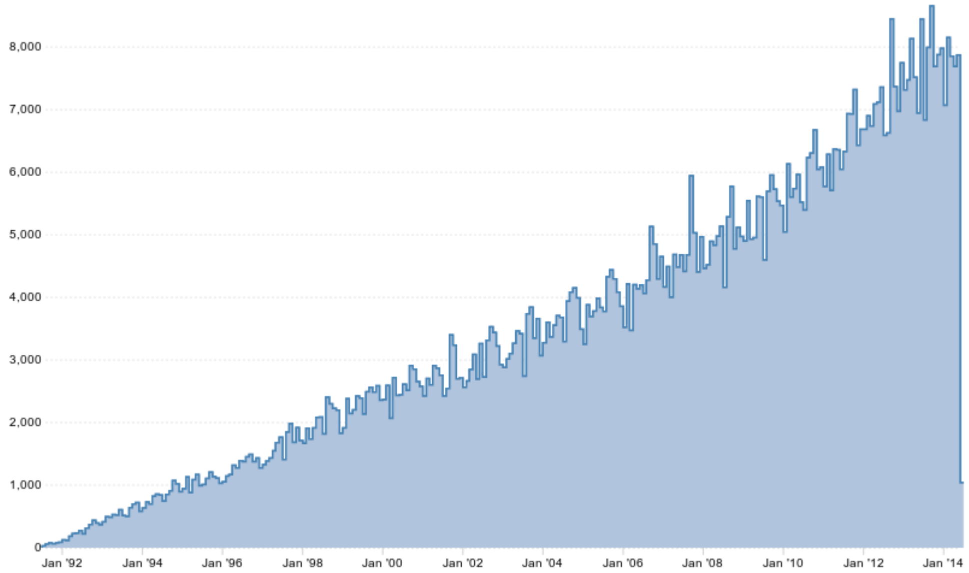


Figure 4.1: Number of new submissions received during each month since August 1991. (From http://arxiv.org/stats/monthly_ submissions)

The total number of retrieved articles up to the date 07.05.2014 is 951,910. The subject of the articles covers physics, mathematics, computer science, nonlinear sciences, quantitative biology and statistics. The submission distribution among different subjects is shown in Figure 4.2.

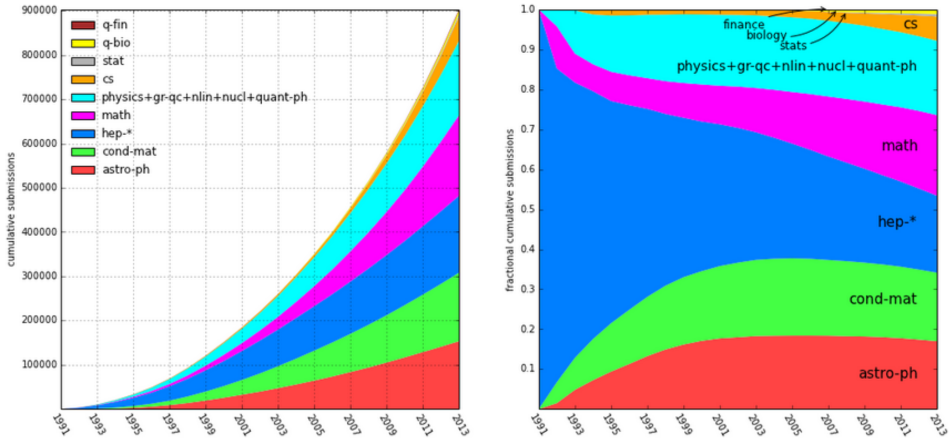


Figure 4.2: The total number of submissions from 1991 to 2013 (left) and percentage of the total submissions (right) for "q-fin" = finance, "biology" = q-bio, "stats" = statistic, "cs" = computer science, "other physics" = physics+nucl (Nuclear)+gr-qc (General Relativity and Quantum Cosmology)+quant-ph (Quantum Physics)+nlin (Nonlinear Sciences), "math" = Mathematics (math+math-ph(Mathematical Physics)), "hep" = High Energy Physics (hep-th(High Energy Physics - Theory)+hep-ph (High Energy Physics - Phenomenology)+hep-lat (High Energy Physics - Lattice)+hep-ex (High Energy Physics - Experiment)), "cond-mat" = Condensed Matter Physics, "astro-ph" = Astrophysics

As we can see from Figure 4.2, articles related to the subject of physics, including "cond-mat (Condensed Matter Physics)", "hep (High Energy Physics)" and other physics occupy over 60 percent of the total submission and the computer science article only take 6 percent. This uneven distribution of subjects, as mentioned in the small-scale experiment , will dramatically hinder the evaluation process since we only have experts in the field of computer science.

To justify the necessity of developing a new interactive exploratory search engine for arXiv, five existing search engines are analysed, including the built-in search engine on the arXiv official web interface,[1] Front of arXiv,[2] arXiv dynamics, [3] SAO/NASA ADS arXiv e-prints Query Form [4] and Paperscape[5].

[1] http://arxiv.org/find

[2] http://front.math.ucdavis.edu/

[3] http://xstructure.inr.ac.ru/

[4] http://adsabs.harvard.edu/preprint_service.html

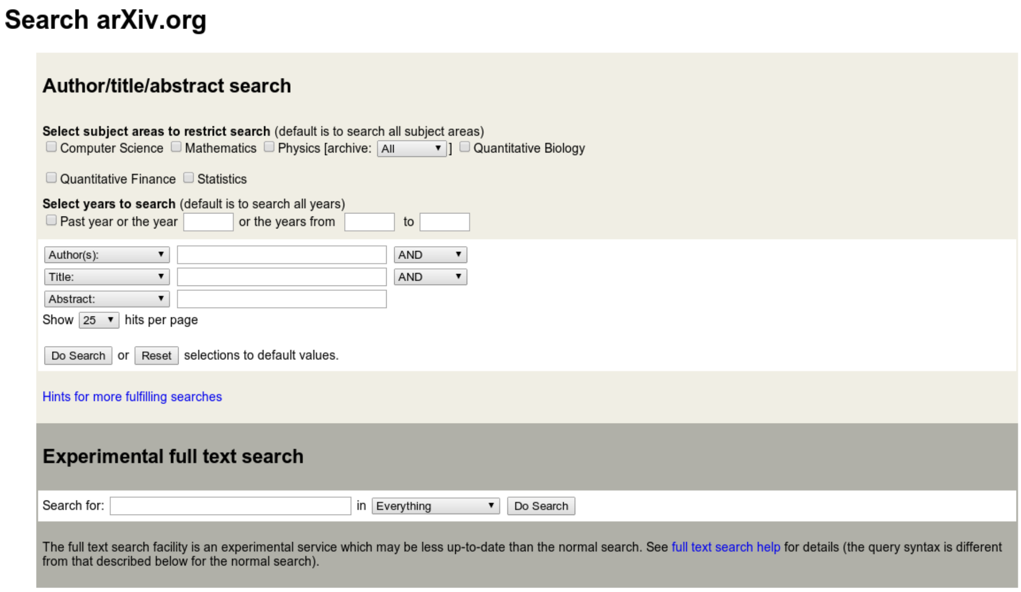
[5] http://paperscape.org/

Their corresponding characteristics are listed below:

- The built-in search engine (shown in Figure 4.3) allows users to search by titles, authors, abstracts, categories and the search can be conducted in full text mode. Particularly, this last mode is unavailable in any other search engines.
- arXiv dynamics (shown in Figure 4.4) provides a search interface similar to the arXiv built-in engine. However, this interface is based on a document classification algorithm. This classification is based on the citation relationship between the papers and classes are organized into a hierarchical scheme. Unfortunately, this classification algorithm is not introduced in detail and no specific paper is indicated.
- Front of arXiv (shown in Figure 4.5) is a search engine maintained by University of California, Davis in United States. It automatically extracts paper abstracts and other metadata from arXiv. Only the searching and browsing functionalities are provided on this interface.
- SAO/NASA ADS arXiv e-prints Query Form (shown in Figure 4.6) is an enhanced search service for arXiv. Besides the common search functions as described for the built-in engine, users could filter the search results by date and sort results by several criteria such as Score, Normalized Score, the publication date, entry date and citation count. The score measures how well each article matches the query. Moreover, the users can decide the weight of fields of authors, title and abstract in the matching. The Normalized Score refers to the score normalized to the number of authors where the articles with fewer authors are preferred.
- Paperscape (shown in Figure 4.7) can be used to search for papers by keywords in title, abstract or authors. This interface is more like a visualization tool than a search engine. All the papers from arXiv are visualised on a map based on the citation relationship. In the visualization, each paper is represented by a circle and the N-body algorithm is used to determine the layout. In this case, each paper is represented as a particle in the N-body problem. Particularly, two forces among the particles are involved in the N-body calculation: “each paper is repelled from all other papers using an anti-gravity inverse-distance force, and each paper is attracted to all of its references using a spring modelled by Hooke’s law”[6].The area of the circle is proportional to the number of citations that paper has, the color indicates the paper’s

[6] http://blog.paperscape.org/?page_id=2

category and the brightness means paper age. When user clicks on a circle, a dialogue box will show up on the screen which contains general information of the selected paper including the title, authors, citation relationship and other details. Users also could create their own citation graph interactively by themselves.



Search arXiv.org

Author/title/abstract search

Select subject areas to restrict search (default is to search all subject areas)
□ Computer Science □ Mathematics □ Physics [archive: All ▾] □ Quantitative Biology

□ Quantitative Finance □ Statistics

Select years to search (default is to search all years)
□ Past year or the year ____ or the years from ____ to ____

Author(s): ▾ ____ AND ▾
Title: ▾ ____ AND ▾
Abstract: ▾ ____
Show 25 ▾ hits per page

Do Search or Reset selections to default values.

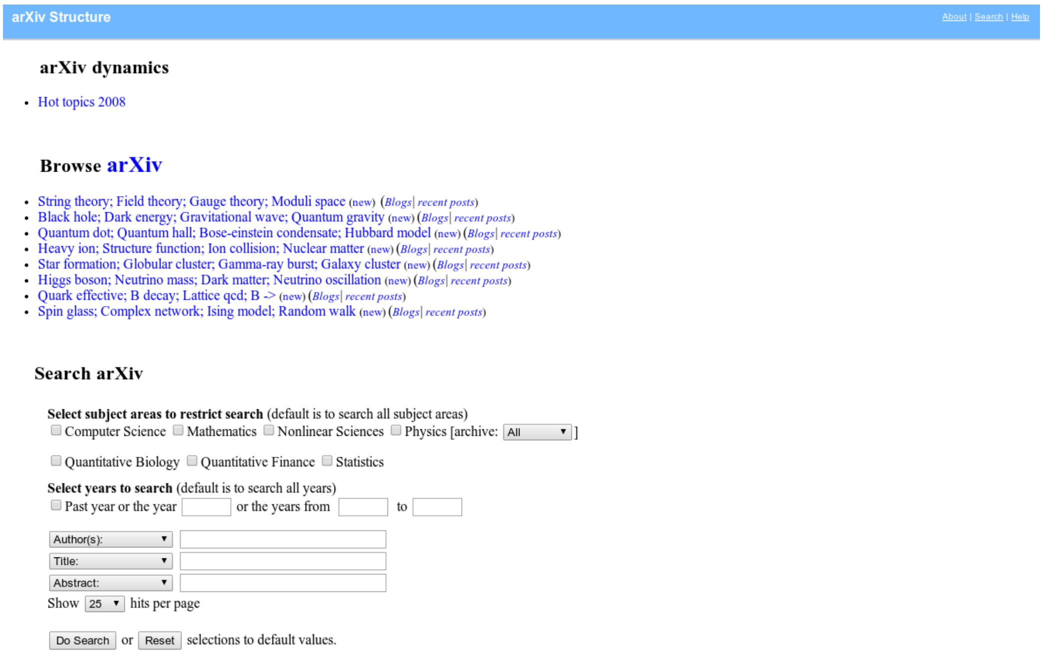
Hints for more fulfilling searches

Experimental full text search

Search for: ____ in Everything ▾ Do Search

The full text search facility is an experimental service which may be less up-to-date than the normal search. See full text search help for details (the query syntax is different from that described below for the normal search).

Figure 4.3: Interface for the built-in search engine



arXiv Structure

arXiv dynamics

- Hot topics 2008

Browse arXiv

- String theory; Field theory; Gauge theory; Moduli space (new) (*Blogs*| *recent posts*)
- Black hole; Dark energy; Gravitational wave; Quantum gravity (new) (*Blogs*| *recent posts*)
- Quantum dot; Quantum hall; Bose-einstein condensate; Hubbard model (new) (*Blogs*| *recent posts*)
- Heavy ion; Structure function; Ion collision; Nuclear matter (new) (*Blogs*| *recent posts*)
- Star formation; Globular cluster; Gamma-ray burst; Galaxy cluster (new) (*Blogs*| *recent posts*)
- Higgs boson; Neutrino mass; Dark matter; Neutrino oscillation (new) (*Blogs*| *recent posts*)
- Quark effective; B decay; Lattice qcd; B -> (new) (*Blogs*| *recent posts*)
- Spin glass; Complex network; Ising model; Random walk (new) (*Blogs*| *recent posts*)

Search arXiv

Select subject areas to restrict search (default is to search all subject areas)
□ Computer Science □ Mathematics □ Nonlinear Sciences □ Physics [archive: All ▾]

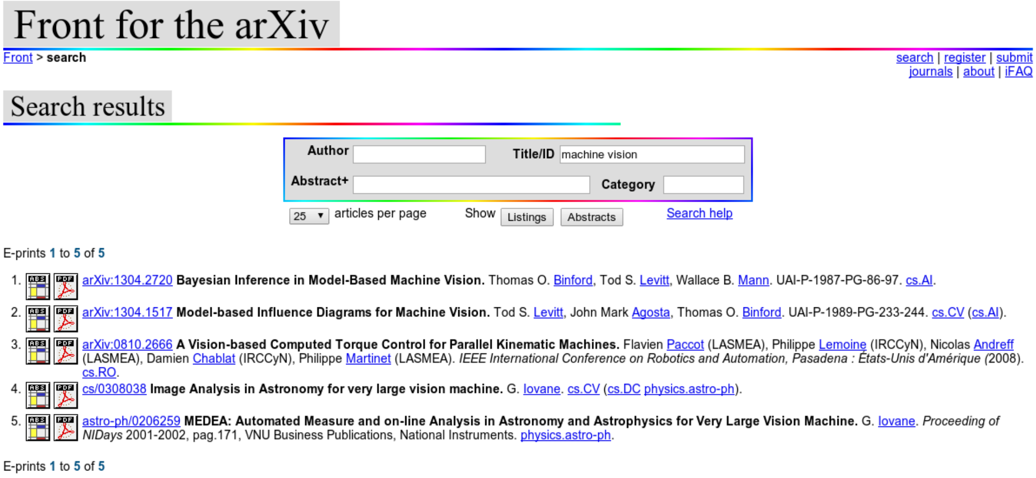
□ Quantitative Biology □ Quantitative Finance □ Statistics

Select years to search (default is to search all years)
□ Past year or the year ____ or the years from ____ to ____

Author(s): ▾ ____
Title: ▾ ____
Abstract: ▾ ____
Show 25 ▾ hits per page

Do Search or Reset selections to default values.

Figure 4.4: Interface for arXiv dynamics



Front for the arXiv

Front > **search**

Search results

Author ____ **Title/ID** machine vision
Abstract+ ____ **Category** ____

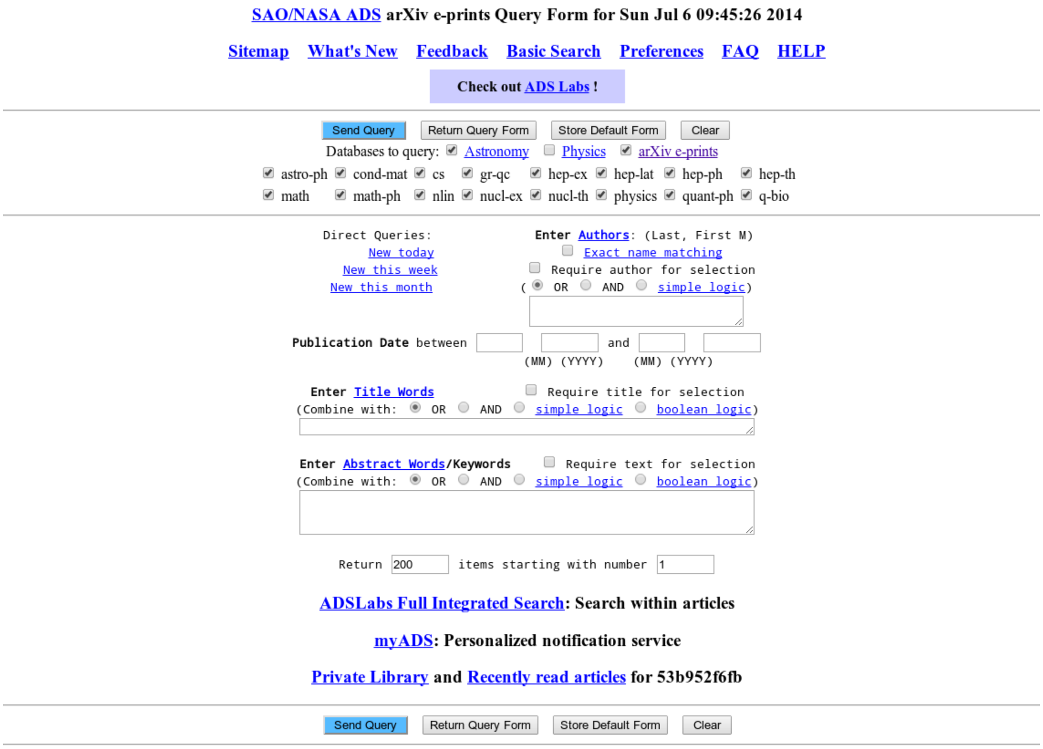
25 ▾ articles per page Show Listings Abstracts Search help

E-prints **1** to **5** of **5**

1. arXiv:1304.2720 **Bayesian Inference in Model-Based Machine Vision.** Thomas O. Binford, Tod S. Levitt, Wallace B. Mann. UAI-P-1987-PG-86-97. cs.AI.
2. arXiv:1304.1517 **Model-based Influence Diagrams for Machine Vision.** Tod S. Levitt, John Mark Agosta, Thomas O. Binford. UAI-P-1989-PG-233-244. cs.CV (cs.AI).
3. arXiv:0810.2666 **A Vision-based Computed Torque Control for Parallel Kinematic Machines.** Flavien Paccot (LASMEA), Philippe Lemoine (IRCCyN), Nicolas Andreff (LASMEA), Damien Chablat (IRCCyN), Philippe Martinet (LASMEA). *IEEE International Conference on Robotics and Automation, Pasadena : États-Unis d'Amérique* (2008). cs.RO.
4. cs/0308038 **Image Analysis in Astronomy for very large vision machine.** G. Iovane. cs.CV (cs.DC physics.astro-ph).
5. astro-ph/0206259 **MEDEA: Automated Measure and on-line Analysis in Astronomy and Astrophysics for Very Large Vision Machine.** G. Iovane. *Proceeding of NIDays* 2001-2002, pag.171, VNU Business Publications, National Instruments. physics.astro-ph.

E-prints **1** to **5** of **5**

Figure 4.5: Interface for Front of arXiv



SAO/NASA ADS arXiv e-prints Query Form for Sun Jul 6 09:45:26 2014

Sitemap What's New Feedback Basic Search Preferences FAQ HELP

Check out ADS Labs !

Send Query | Return Query Form | Store Default Form | Clear

Databases to query: ☑ Astronomy □ Physics ☑ arXiv e-prints

☑ astro-ph ☑ cond-mat ☑ cs ☑ gr-qc ☑ hep-ex ☑ hep-lat ☑ hep-ph ☑ hep-th
☑ math ☑ math-ph ☑ nlin ☑ nucl-ex ☑ nucl-th ☑ physics ☑ quant-ph ☑ q-bio

Direct Queries:
New today
New this week
New this month

Enter Authors: (Last, First M)
□ Exact name matching
□ Require author for selection
(● OR ○ AND ○ simple logic)

Publication Date between ______ ______ and ______ ______
(MM) (YYYY) (MM) (YYYY)

Enter Title Words □ Require title for selection
(Combine with: ● OR ○ AND ○ simple logic ○ boolean logic)

Enter Abstract Words/Keywords □ Require text for selection
(Combine with: ● OR ○ AND ○ simple logic ○ boolean logic)

Return 200 items starting with number 1

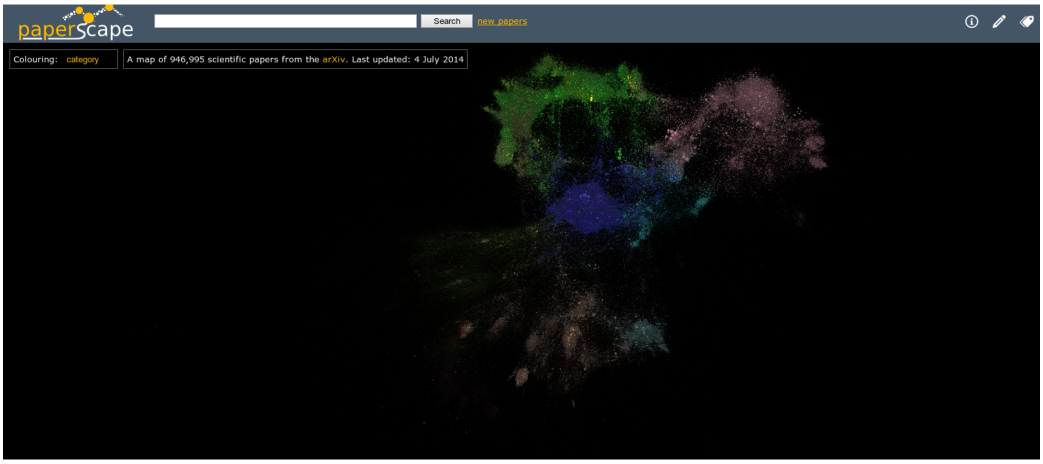
ADSLabs Full Integrated Search: Search within articles

myADS: Personalized notification service

Private Library and Recently read articles for 53b952f6fb

Send Query | Return Query Form | Store Default Form | Clear

Figure 4.6: Interface for SAO/NASA ADS arXiv e-prints Query Form



Colouring: category

A map of 946,995 scientific papers from the arXiv. Last updated: 4 July 2014

Figure 4.7: Interface for Paperscape

Based on the characteristics listed above, we can observe that most of the existing search engines are merely designed to meet the basic search needs, although some enhancements are provided (e.g. SAO/NASA ADS arXiv e-prints Query Form allows users to sort the search results by different criteria). Interactive visualization is provided in Paperscape. However, it is only dedicated for the citation relationship not for paper contents.

We also find other interesting web services for arXiv, including the "bookworm"[7] which analyses and presents research trends within the papers in arXiv and a demo website[8] for a ranking algorithm for papers in arXiv. However, these works only have a weak connection to arXiv from the point of view as a search engine.

Consequently, we can claim that currently no existing arXiv search engines could support exploratory search and our contribution will improve this situation dramatically.

4.2 Interactive Exploratory Search System for ArXiv

In this thesis, an interactive exploratory search system (IESS) is built for the online scientific article database, arXiv. With this system, users can search the arXiv repository by typing queries as usual but can then direct their search in a novel way by interacting with keywords. The highlight of the system is the users' intent visualization and the user-system interaction. Specially, the users' search intents are estimated on an Intent Radar layout and users can interact with this search system by manipulating the estimated intents. The intents are estimated by the interactive intent modelling, as introduced in Chapter 2 [9]. In some way, this IESS for arXiv can be considered an extension of the system prototype proposed by the paper by Ruotsalo et al. [17]. However, compared with the system proposed in the paper, a new part of the user interface, the workspace, is introduced in the thesis and the rest of the interface is still the same as in the paper [17]. The other differences are the underlying database and the keywords that are extracted by the AKE algorithm. In the original paper, the keywords are provided by the authors. In the following, the IESS will be described mainly from two aspects: the system user interface and the underlying arXiv data set.

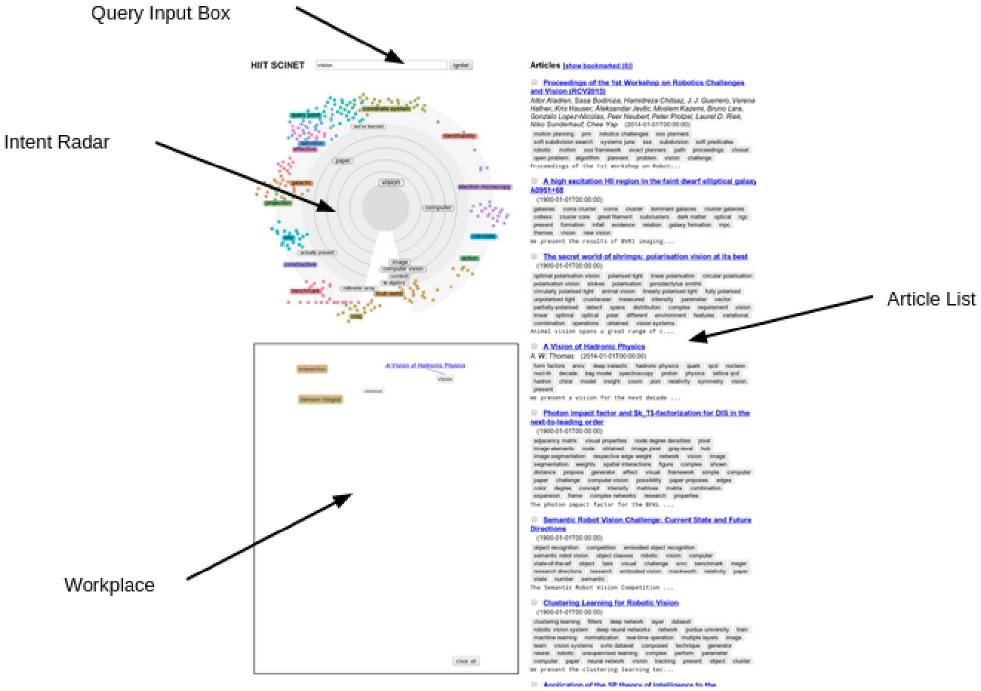
[7] http://bookworm.culturomics.org/arxiv/

[8] http://arxivsorter.org/

[9] More detailed information about the interactive intent modelling can be found in Chapter 2 and the paper by Ruotsalo et al. [17] and it is not discussed in this part of the thesis.

4.2.1 System User Interface

The system user interface consists of four components, including a query input box, an Intent Radar, a workspace and an article list. The layout of the interface is shown in Figure 4.8.



Articles [show bookmarked (0)]

□ Proceedings of the 1st Workshop on Robotics Challenges and Vision (RCV2013)
Aitor Aladren, Sasa Bodiroza, Hamidreza Chitsaz, J. J. Guerrero, Verena Hafner, Kris Hauser, Aleksandar Jevtic, Moslem Kazemi, Bruno Lara, Gonzalo Lopez-Nicolas, Peer Neubert, Peter Protzel, Laurel D. Riek, Niko Sunderhauf, Chee Yap (2014-01-01T00:00:00)
motion planning pm robotics challenges sss planners soft subdivision search systems june sss subdivision soft predicates robotic motion sss framework exact planners path proceedings closed open problem algorithm planners problem vision challenge
Proceedings of the 1st Workshop on Robot...

□ A high excitation HII region in the faint dwarf elliptical galaxy A0951+68
(1900-01-01T00:00:00)
galaxies coma cluster coma cluster dominant galaxies cluster galaxies colless cluster core great filament subclusters dark matter optical ngc present formation infall evidence relation galaxy formation ngc themes vision new vision
We present the results of BVRI imaging...

□ The secret world of shrimps: polarisation vision at its best
(1900-01-01T00:00:00)
optimal polarisation vision polarised light linear polarisation circular polarisation polarisation vision stokes polarisation gonodactylus smithii circularly polarised light animal vision linearly polarised light fully polarised unpolarised light crustacean measured intensity parameter vector partially-polarised detect spans distribution complex requirement vision linear optimal optical polar different environment features variational combination operations obtained vision systems
Animal vision spans a great range of c...

□ A Vision of Hadronic Physics
A. W. Thomas (2014-01-01T00:00:00)
form factors arxiv deep inelastic hadronic physics quark qcd nucleon nucl-th decade bag model spectroscopy proton physics lattice qcd hadron chiral model insight cosm pion relativity symmetry vision present
We present a vision for the next decade ...

□ Photon impact factor and k_T-factorization for DIS in the next-to-leading order
(1900-01-01T00:00:00)
adjacency matrix visual properties node degree densities pixel image elements node obtained image pixel grey-level hub image segmentation respective edge weight network vision image segmentation weights spatial interactions figure complex shown distance propose generator effect visual framework simple computer paper challenge computer vision possibility paper proposes edges color degree concept intensity matrices matrix combination expansion frame complex networks research properties
The photon impact factor for the BFKL ...

□ Semantic Robot Vision Challenge: Current State and Future Directions
(1900-01-01T00:00:00)
object recognition competition embodied object recognition semantic robot vision object classes robotic vision computer state-of-the-art object task visual challenge srvc benchmark major research directions research embodied vision mackworth relativity paper state number semantic
The Semantic Robot Vision Competition ...

□ Clustering Learning for Robotic Vision
(1900-01-01T00:00:00)
clustering learning filters deep network layer dataset robotic vision system deep neural networks network purdue university train machine learning normalization real-time operation multiple layers image team vision systems svhn dataset composed technique generator neural robotic unsupervised learning complex perform parameter computer paper neural network vision tracking present object cluster
We present the clustering learning tec...

□ Application of the SP theory of intelligence to the

Figure 4.8: A interface for the IESS for arXiv

Users can submit their queries through the query input box. Similar to other search engines, queries are in the form of single or combined keywords. After a query is issued into the search system, the user's intents are estimated and visualized on the Intent Radar. Then the user will direct her exploration by manipulating the keywords on the Intent Radar.

Particularly, two kinds of search intents are estimated, including the current search intents and the future search intents. Current search intents are the direct estimation of the user's intents and located within the inner circles around the solid center. The distance between the estimated intents and the center indicates their relevance to the current estimated search intent. The angles of different keywords indicate the similarity between the intents. The future intents are visualized in the outer circles. These future intents are defined as the potential search directions that the users would follow if they are interested in the corresponding current intents within the inner circle. Due to the limited space of the radar layout, most of the future intents are designed to be observed by the fisheye lens.

Users can give their feedback to the estimated intents by dragging the keyword closer or further to the center as shown in Figure 4.9. After the user moves the intents, the search results would be updated accordingly.

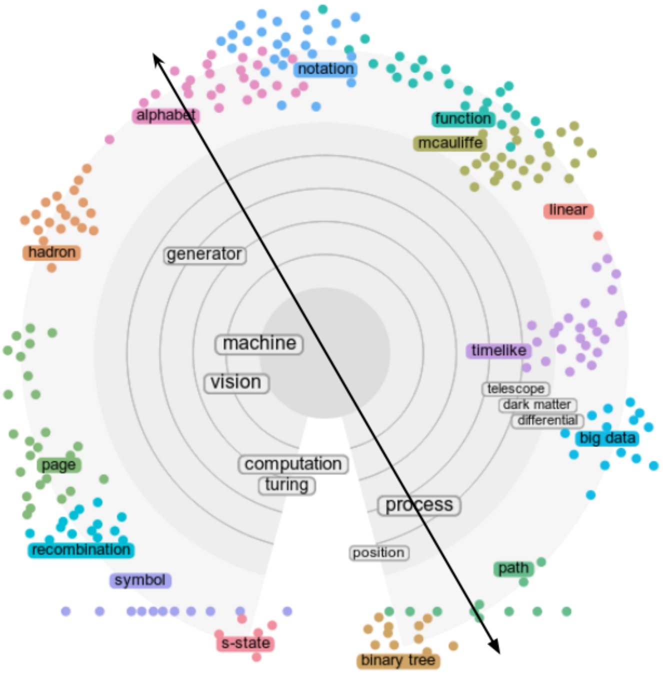
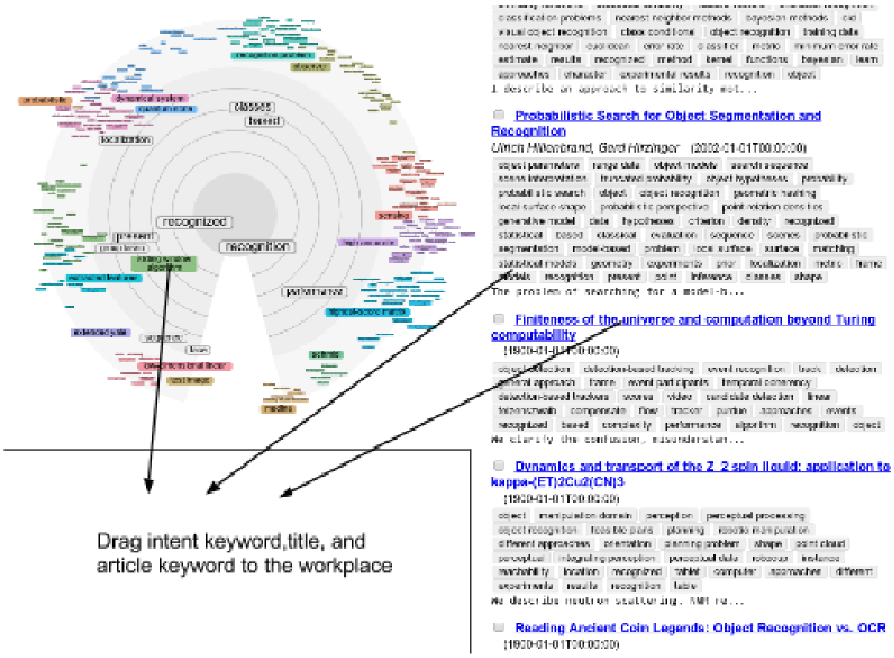


Figure 4.9: keywords can be dragged on the Intent Radar along the straight line between the keyword and the circle center

Right below the Intent Radar is the workspace, which is designed as a mind map. Users can drag keywords and document titles onto this workplace as shown in Figure 4.10. Keywords can also be dragged into the workspace both from the radar and from the keywords shown within the document list. These dragged items can be placed and organized freely within the workspace. Moreover, the title of a document and the keywords will be automatically linked together if they belong to the same document, as shown in Figure 4.11. Particularly, a "clear all" button is located at the bottom of the workspace. Users can click the button to remove all the information within the workspace. The workspace is created for two purposes. First, the workspace can be used as a notebook where users can record all the search findings during the search session. Second, the workspace can be used as a tool to structure the search finding since the dragged item can be placed in arbitrary locations. In this way, the workspace will help the participant to better remember and understand the search results.



I describe an approach to similarity met...

□ Probabilistic Search for Object Segmentation and Recognition
Ulrich Hillenbrand, Gerd Hirzinger (2002-01-01T00:00:00)

The problem of searching for a model-b...

□ Finiteness of the universe and computation beyond Turing computability
(1900-01-01T00:00:00)

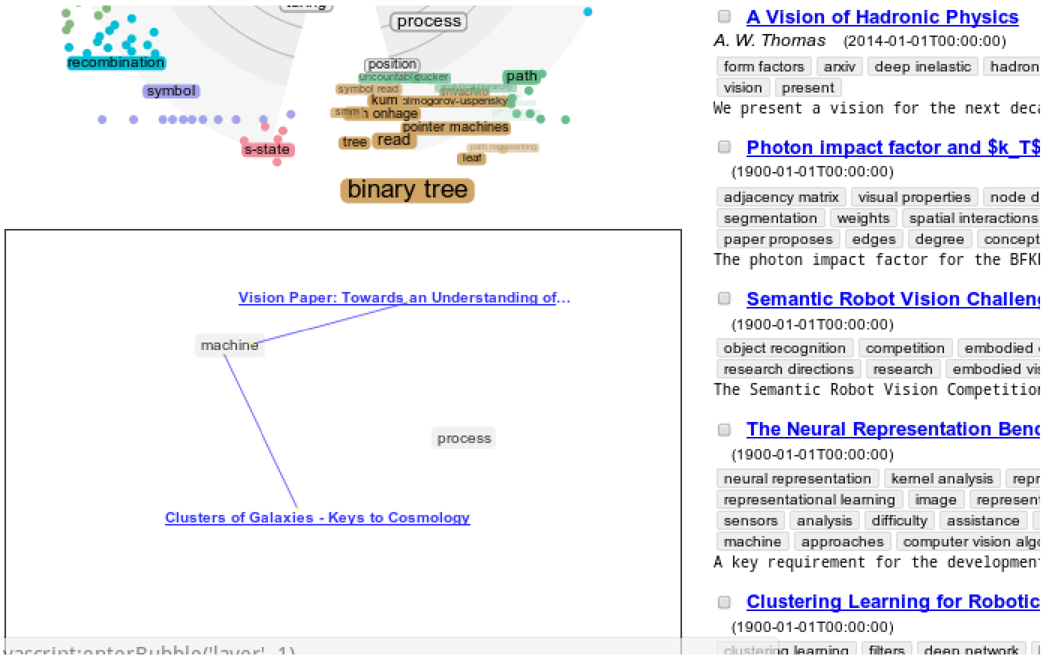
We clarify the confusion, misunderstan...

□ Dynamics and transport of the Z_2 spin liquid: application to kappa-(ET)2Cu2(CN)3
(1900-01-01T00:00:00)

We describe neutron scattering, NMR re...

□ Reading Ancient Coin Legends: Object Recognition vs. OCR
(1900-01-01T00:00:00)

Figure 4.10: users can drag keywords on the Intent Radar and keywords and title within the article list area to the workspace.



□ A Vision of Hadronic Physics
A. W. Thomas (2014-01-01T00:00:00)
form factors | arxiv | deep inelastic | hadron
vision | present
We present a vision for the next deca

□ Photon impact factor and k_T
(1900-01-01T00:00:00)
adjacency matrix | visual properties | node d
segmentation | weights | spatial interactions
paper proposes | edges | degree | concept
The photon impact factor for the BFKI

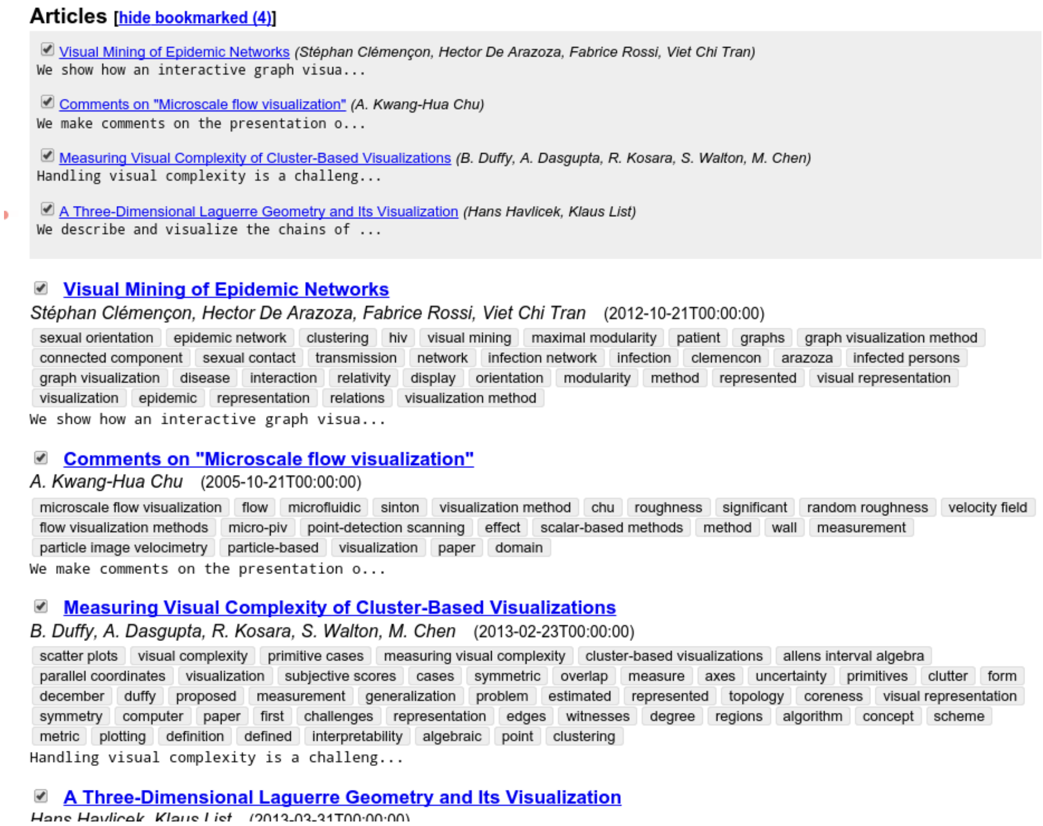
□ Semantic Robot Vision Challen
(1900-01-01T00:00:00)
object recognition | competition | embodied
research directions | research | embodied vi
The Semantic Robot Vision Competitio

□ The Neural Representation Benc
(1900-01-01T00:00:00)
neural representation | kernel analysis | repr
representational learning | image | represen
sensors | analysis | difficulty | assistance
machine | approaches | computer vision alg
A key requirement for the developmen

□ Clustering Learning for Robotic
(1900-01-01T00:00:00)
clustering learning | filters | deep network

Figure 4.11: keywords and titles within the workspace will be linked automatically if they belong to the same document

On the right side of the user interface is the document list where users can bookmark articles, read their abstracts and keywords and follow the link to find the full-context paper. Specially, the bookmarked articles will be stored through the whole search session and shown on the top of the document list as shown in Figure 4.12.



Articles [hide bookmarked (4)]

☑ Visual Mining of Epidemic Networks *(Stéphan Clémençon, Hector De Arazoza, Fabrice Rossi, Viet Chi Tran)*
We show how an interactive graph visua...

☑ Comments on "Microscale flow visualization" *(A. Kwang-Hua Chu)*
We make comments on the presentation o...

☑ Measuring Visual Complexity of Cluster-Based Visualizations *(B. Duffy, A. Dasgupta, R. Kosara, S. Walton, M. Chen)*
Handling visual complexity is a challeng...

☑ A Three-Dimensional Laguerre Geometry and Its Visualization *(Hans Havlicek, Klaus List)*
We describe and visualize the chains of ...

☑ **Visual Mining of Epidemic Networks**
Stéphan Clémençon, Hector De Arazoza, Fabrice Rossi, Viet Chi Tran (2012-10-21T00:00:00)
sexual orientation | epidemic network | clustering | hiv | visual mining | maximal modularity | patient | graphs | graph visualization method | connected component | sexual contact | transmission | network | infection network | infection | clemencon | arazoza | infected persons | graph visualization | disease | interaction | relativity | display | orientation | modularity | method | represented | visual representation | visualization | epidemic | representation | relations | visualization method
We show how an interactive graph visua...

☑ **Comments on "Microscale flow visualization"**
A. Kwang-Hua Chu (2005-10-21T00:00:00)
microscale flow visualization | flow | microfluidic | sinton | visualization method | chu | roughness | significant | random roughness | velocity field | flow visualization methods | micro-piv | point-detection scanning | effect | scalar-based methods | method | wall | measurement | particle image velocimetry | particle-based | visualization | paper | domain
We make comments on the presentation o...

☑ **Measuring Visual Complexity of Cluster-Based Visualizations**
B. Duffy, A. Dasgupta, R. Kosara, S. Walton, M. Chen (2013-02-23T00:00:00)
scatter plots | visual complexity | primitive cases | measuring visual complexity | cluster-based visualizations | allens interval algebra | parallel coordinates | visualization | subjective scores | cases | symmetric | overlap | measure | axes | uncertainty | primitives | clutter | form | december | duffy | proposed | measurement | generalization | problem | estimated | represented | topology | coreness | visual representation | symmetry | computer | paper | first | challenges | representation | edges | witnesses | degree | regions | algorithm | concept | scheme | metric | plotting | definition | defined | interpretability | algebraic | point | clustering
Handling visual complexity is a challeng...

☑ **A Three-Dimensional Laguerre Geometry and Its Visualization**
Hans Havlicek, Klaus List (2013-03-31T00:00:00)

Figure 4.12: marked articles are stored on the top of the document list.

4.2.2 Data Set Acquisition and Data Processing

A series of data processing steps are conducted besides the design of the user interface, as shown in Figure 4.13

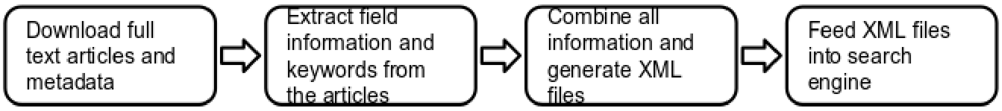

Download full text articles and metadata
Extract field information and keywords from the articles
Combine all information and generate XML files
Feed XML files into search engine

Figure 4.13: data processing flow chart

All the arXiv data is downloaded through the arXiv bulk data access APIs since massively crawling on the arXiv website is forbidden. In detail, arXiv provides a set of APIs to support real-time programmatic access to the article metadata[10]. As for the full-text access, the full-text contents are stored on the Amazon Simple Storage Service (Amazon S3) and thus can be obtained from it. In total, 905,507 documents up to the date 14.01.2014 are collected. The data includes the full text document and the metadata. The full text documents are in the PDF format. In order to process the content, these PDF files are first converted into TXT format. The metadata provides the following necessary information for the search system: title, author names, abstract and the publication year. Since keywords cannot be found in the metadata, they will be obtained from the full article text. The keywords generated from the full article text and all the other field information will be assembled into XML files and fed into the search system for indexing.

[10] Metadata refers to the descriptive information about each article in arXiv. The metadata for each article contains the information of title, abstract, submit date, authors, subjects, the DOI, the journal, the URL and the identifier.

Chapter 5

User Experiment on the Search System for arXiv

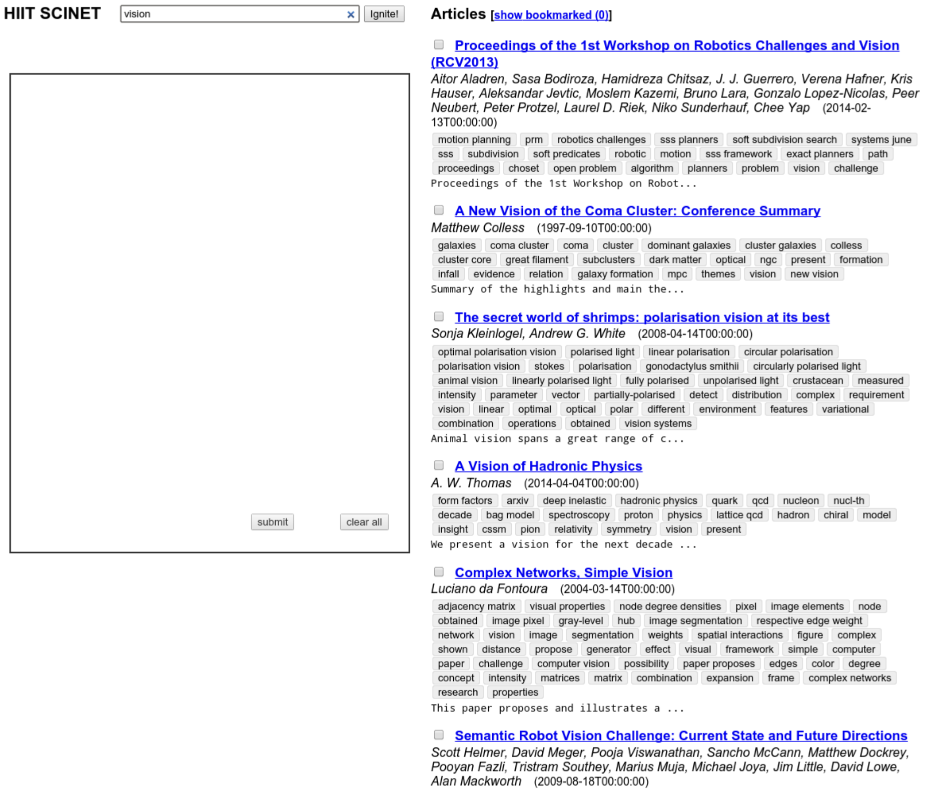
In this chapter, a small-scale user experiment is conducted to verify the usability of the interactive exploratory search system (IESS) for arXiv [1]. Similar to the experiments in the paper by Ruotsalo et al. [17], this experiment is designed in a 2×2 fashion with two search tasks and two system setups. Each participant is required to complete these two search tasks with two variant search systems on a provided computer. Each participant performs each task on one system only and the order of which system is used for which task is alternated between participants. With this experiment, we want to investigate whether this IESS can help the users with the exploratory search and achieve better task performance.

5.1 Variant Systems

Two systems are used in the experiment, including the designed IESS and a variant search system without the Intent Radar as shown in Figure 5.1. The removal of the Intent Radar is aimed to test its effect to the search-supporting functionalities. Specially, a submit button, as shown in Figure 5.1, is added to the workspace for the experiment, allowing participants to submit the logs, which will be introduced in the section of Evaluation and Measurement. Participants will be instructed about how to use the systems at the beginning of the user experiment. This settings is similar to the original paper by Ruotsalo et al. [17]. However, the keywords are extracted by the AKE algorithm and the new workspace interface component is added,

[1] This experiment is intended as a small feasibility study and a larger study will be conducted in Chapter 6.

which makes the experiment setting different from the original one.



Articles [show bookmarked (0)]

□ Proceedings of the 1st Workshop on Robotics Challenges and Vision (RCV2013)
Aitor Aladren, Sasa Bodiroza, Hamidreza Chitsaz, J. J. Guerrero, Verena Hafner, Kris Hauser, Aleksandar Jevtic, Moslem Kazemi, Bruno Lara, Gonzalo Lopez-Nicolas, Peer Neubert, Peter Protzel, Laurel D. Riek, Niko Sunderhauf, Chee Yap (2014-02-13T00:00:00)
motion planning prm robotics challenges sss planners soft subdivision search systems june sss subdivision soft predicates robotic motion sss framework exact planners path proceedings choset open problem algorithm planners problem vision challenge
`Proceedings of the 1st Workshop on Robot...`

□ A New Vision of the Coma Cluster: Conference Summary
Matthew Colless (1997-09-10T00:00:00)
galaxies coma cluster coma cluster dominant galaxies cluster galaxies colless cluster core great filament subclusters dark matter optical ngc present formation infall evidence relation galaxy formation mpc themes vision new vision
`Summary of the highlights and main the...`

□ The secret world of shrimps: polarisation vision at its best
Sonja Kleinlogel, Andrew G. White (2008-04-14T00:00:00)
optimal polarisation vision polarised light linear polarisation circular polarisation polarisation vision stokes polarisation gonodactylus smithii circularly polarised light animal vision linearly polarised light fully polarised unpolarised light crustacean measured intensity parameter vector partially-polarised detect distribution complex requirement vision linear optimal optical polar different environment features variational combination operations obtained vision systems
`Animal vision spans a great range of c...`

□ A Vision of Hadronic Physics
A. W. Thomas (2014-04-04T00:00:00)
form factors arxiv deep inelastic hadronic physics quark qcd nucleon nucl-th decade bag model spectroscopy proton physics lattice qcd hadron chiral model insight cssm pion relativity symmetry vision present
`We present a vision for the next decade ...`

□ Complex Networks, Simple Vision
Luciano da Fontoura (2004-03-14T00:00:00)
adjacency matrix visual properties node degree densities pixel image elements node obtained image pixel gray-level hub image segmentation respective edge weight network vision image segmentation weights spatial interactions figure complex shown distance propose generator effect visual framework simple computer paper challenge computer vision possibility paper proposes edges color degree concept intensity matrices matrix combination expansion frame complex networks research properties
`This paper proposes and illustrates a ...`

□ Semantic Robot Vision Challenge: Current State and Future Directions
Scott Helmer, David Meger, Pooja Viswanathan, Sancho McCann, Matthew Dockrey, Pooyan Fazli, Tristram Southey, Marius Muja, Michael Joya, Jim Little, David Lowe, Alan Mackworth (2009-08-18T00:00:00)

Figure 5.1: the variant search system without the Intent Radar

5.2 Search Tasks

The tasks are designed as information seeking tasks. Each task consists of 5 questions in a specific field. Participants are required to answer these questions by searching information with the IESS systems and write down the answers in an online answer sheet. The task fields are chosen as “computer vision” and “text mining”. Their corresponding questions are listed in Table 5.1 and Table 5.2 below.

Table 5.1: questions for the field: computer vision

1	List at least 3 methods or algorithms for image feature extraction
2	List at least 3 methods or algorithms for image segmentation
3	List at least 3 sub-tasks for texture analysis
4	List at least 3 algorithms or techniques for object detection
5	List at least 3 algorithms for image compression

Table 5.2: questions for the field: text mining

1	Please list at least three application scenarios for text analysis
2	Please list at least three algorithms for text clustering
3	Please list at lease three application areas for sentiment analysis
4	Please list at least three algorithms of spam detection related to Internet applications
5	Please list two application scenarios for record linkage

5.3 Participants

Two Participants are recruited from the Department of Information and Computer Science of Aalto University. They are graduate students with a study background in computer science and both of them have the experience of literature searching before. A prior knowledge survey is conducted at the start of the experiment by asking the participants to fill in a questionnaire, in which the participants will evaluate their knowledge on different fields. Based on the survey, the participants are neither experts nor novices on these fields.

5.4 Evaluation and Measurement

The task performance is measured by the quality of the participants' answers and evaluated based on their accuracy and completeness. The score for each question ranges from 1 to 3 points and the maximum score for each task is $3 \times 5 = 15$ points.

When the participants are using the search systems, all the interactions between the participants and the systems are logged in a text file. The types of logged interactions are listed in Table 5.3.

Table 5.3: list of logged interactions

1	Drag a keyword within the Intent Radar
2	Drag a keyword on the radar to the panel
3	Drag a title to the panel
4	Drag a keyword under an article title to the panel
5	Delete an article or a keyword on the workplace
6	Bookmark an article
7	Unbookmark an article
8	Send a search request from the query box
9	Send a request for update based on dragging of keywords on the Intent Radar
10	Click to see an abstract
11	Click to close an abstract
12	Click to a link to see the original paper
13	Click clear button to clear the workspace

5.5 Results and Conclusion

Based on the accuracy and completeness of the answers, the scores of two tasks are calculated and listed in Table 5.4, in which CV denotes computer vision , TM denotes text mining and the scores are normalized by dividing them by the maximum score 15.

Table 5.4: scores for both tasks on variant systems

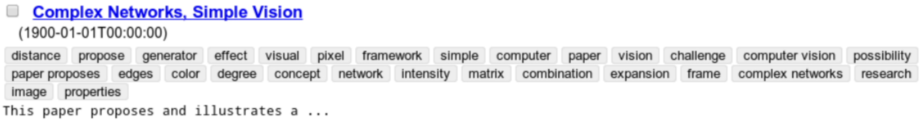
	task 1 (CV)	task 2 (TM)	average score
system with Intent Radar	0.86	0.67	0.77
system without Intent Radar	0.73	0.73	0.73

As shown in Table 5.4, participants achieve better performance on the system with the Intent Radar.

More interestingly, the answers for the same task are totally different, which suggests that the Intent Radar performs as an important information source.

Considering the slightly higher task scores with the Intent Radar, we can conclude that this interactive exploratory search system for arXiv may positively contribute to the information exploration tasks.

On the other hand, we also detected two problems about the system based on the experiment and the user feedback. The first problem is that too many keywords are listed under the title, which may hinder the user from searching for information. The next problem is that the quality of the keywords is not satisfactory and many common words are recognized as keywords. As a result, the keywords shown on the Intent Radar are less informative than the ones in the original paper. For example, as shown in Figure 5.2, the extracted keywords for the article "Complex Networks, Simple Vision" include some common words, such as distance, effect and pixel, which should not be recognised as keywords. Obviously, this IESS for arXiv still has room for improvement.



□ Complex Networks, Simple Vision
(1900-01-01T00:00:00)
distance propose generator effect visual pixel framework simple computer paper vision challenge computer vision possibility paper proposes edges color degree concept network intensity matrix combination expansion frame complex networks research image properties
This paper proposes and illustrates a ...

Figure 5.2: keywords of an article in arXiv extracted by KP-Miner

This brief test adds to the related experiment in the paper by Ruotsalo et al. [17] regarding the usefulness of Intent Radar but is itself only a feasibility study. However, it is so far the only study of Intent Radar on arXiv. Further larger tests are needed for conclusive results.

Chapter 6

Sensemaking Ability of the Scinet Search System

In this chapter, we make the hypothesis that the interactive exploratory search engine is able to assist with sensemaking due to the relationship between the keywords on the Intent Radar. In the following, we will review the relevant work about sensemaking and a user experiment will be conducted to investigate this sensemaking effect.

6.1 Related Work

Sensemaking is a hot research topic in the field of intelligent systems (see, for example, [10], [13]) and closely related to the study of learning process and cognitive science (see, for example, [2],[24],[13]). In the past, sensemaking has been studied from different viewpoints.

In the study of Klein et al.[10], the authors have discussed the role of sensemaking for intelligent systems. Specifically, according to Klein et al., sensemaking studies are studies from three perspectives, including psychology, human-centred computing and the perspective of naturalistic decision making. They did not provide a universal definition of the meaning of sensemaking but reviewed and refuted some sensemaking theories. For example, from the perspective of psychology, the authors deem that sensemaking is related to creativity, comprehension, curiosity, mental modeling, explanation and situational awareness. However, sensemaking is essentially different from these factors. Sensemaking should be a "motivated continuous effort to understand connections" among people, places and events.

In the paper by Pirolli and Russell [14], the authors claim that sensemaking is not equal to information retrieval. Instead, they describe sensemaking as “an active processing of information to achieve understanding”. In their opinion, sensemaking is related to learning about “new domains, solving ill-structured problems, acquiring situational awareness and participating in social exchanges of knowledge”. They also proposed another three perspectives to consider sensemaking, including the representation construction model of sensemaking, where the sensemaking process is organized into two main loops of activities, involving the a foraging loop and a sensemaking loop, the data/frame perspective of sensemaking, which regards the sensemaking process as achieving a “mental model representation of the state of the affairs in the world” [14] and collaborative sensemaking, where a group of people work together to make sense of the information they are holding. These perspectives are different from the perspectives mentioned in Klein et al. [10]. However, they can be considered a subdivision for the perspective of human-centered computing.

In the essay from Abraham et al. [1], the authors have studied the interaction between experienced information processors and the sensemaking process for the web-based information sources since they realise that there is a need to develop tools to support sensemaking for the everyday web-based practices. They regard sensemaking as “the strategies and behaviours evident when users collect, evaluate, understand, interpret, and integrate new information for their own specific problem/task needs”. Further, the process of information interaction is also divided into four categories, including search, evaluation for selection, evaluation for use and use.

In the paper by Fisher et al. [7], distributed sensemaking is studied where users’ sensemaking tasks are assisted by a previous sensemaking work from anonymous users in a collaborative-like distributed setting. The authors deem that sensemaking is used to construct a mental representation of interrelated pieces of information to accomplish a task.

In our work, we will consider sensemaking as a learning process, or more precisely, a comprehension process. We deem that when users are searching information through a search system, retrieving the needed information is only the first step. After that, users need to further understand the search results. More ideally, they can learn new knowledge from the results or obtain a deeper understanding the search topics. In the following experiment, we will use topic comprehension tasks to examine whether the interactive exploratory search system will support sensemaking.

6.2 Sensemaking User Experiment

In this user experiment, we want to investigate whether the interactive exploratory search system has the ability to help users' sensemaking. The user experiment is designed with a topic comprehension setting, involving two variant systems and 24 general topics. The topics are mainly from the computer science field but topics about physics, chemistry, ecology, and social sciences are also included. In the experiment, each participant will be asked to explore eight topic domains which are selected from those 24 topics and find main concepts and sub-concepts related to the general topics.

The design of the user experiment is guided by the suggestions proposed in the work by Wildemuth and Freund [23]. The suggestions are valuable for our work although they are initially aimed for eliciting exploratory search behaviours. According to their research, the tasks should be general, ambiguous, ill-structured and complex enough in order to elicit exploratory behaviours. Our experiment design is consistent with these suggestions.

6.2.1 Tasks

The tasks used in the user experiment are about topic comprehension. The participants need to conduct eight comprehension tasks based on two search systems, four tasks for each system. Each comprehension task is concerning a general topic. For each general topic, participants are required to describe at least three related main concepts and describe as many as possible sub-concepts related to each of the main concepts. Suppose the overall topic is "computer vision", then the participants need to find at least three main concepts related to computer vision, such as "image processing", "3D representation" and "robotics" . After that, sub-concepts related to each found main concept should be found. For "3D representation", the sub-concepts can be "Marr's theory" and "visual content".

The eight general topics used in the participant experiment are selected from a topic pool, which consists of 24 general topics. Most of the topics are from the computer science domain. However, other domains including physics, chemistry, ecology, and social sciences are included as well. Those 24 topics are listed in Table 6.1.

Table 6.1: 24 overall topics in the experiment

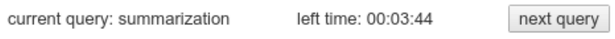
computer vision	cryptography	natural language processing
web search	distributed system	compiler design
human memory	image compression	dimension reduction
graphical model	kernel function	wearable sensors
web design	combinatorial optimization	digital content sharing
biodiversity	synthetic chemistry	molecular mechanics
3D graphics	cognitive psychology	information visualization
sensor network	information retrieval evaluation	human computer interaction

The experiment is designed to test the sensemaking ability of the IESS instead of the ability to direct search which was tested in the paper by Ruotsalo et al. [17]. Particularly, search queries, which are the general topics themselves, are pre-specified and the queries are automatically issued into the search systems. Users will be given a brief time to inspect each result set, simulating the brief time that a user of an interactive search system would spend on an individual result set before continuing the search. Each task has to be completed in five minutes. For the first two minutes, participants are required to read and collect information on the screen by dragging the useful items to the workspace. After that, all the information on the screen will disappear except the workspace. Participants can only use the information shown on the interface and are not allowed to perform any additional searches outside the IESS that they are using. Next, participants will have three minutes to write the answers. Particularly, they will write the answers by checking their notes on the workspace, which is intended as a memory aid based on what they comprehend as important in this experiment .

6.2.2 Variant Systems

Two variant search systems are used in the experiment, including the complete version of the IESS and the baseline version of the IESS without the Intent Radar. In the complete version of the system, participants are expected to use both the Intent Radar and the article list to gather information. In the baseline system without the Intent Radar, participants are only allowed to utilize the article list. We hypothesize that users' sensemaking ability is increased when they see the relationships between the keywords on the Intent Radar. The keywords are intended as the estimation of the search intents, which can help users enlarge the information space. Moreover, the similarities among the keywords are also indicated by locations and angles. Thus the information on the Intent Radar is supposed to increase the users' ability of senesemaking and participants who use the full version of the search system are expected to achieve better performance in the user experiment.

Compared with the IESS system described in Chapter 4, the system interfaces are modified according to the setting of the experiment. Firstly, the search box is removed and replaced with an information panel. The information panel consists of three parts. The first part shows the current topic. In the second part, a count-down timer is created, which indicates the time left for the current topic. Next to the timer is the jump-to-next button, with which the participant can choose to jump to next topic directly if s/he completes one topic earlier. This information panel is shown in Figure 6.1. Secondly, compared to the Scinet IESS in [17], our interface also includes the workspace, described in Chapter 4. However, here the purpose of the addition is not to investigate effects of having a workspace in the system. Instead, the workspace is used as a tool to gather more information about users' sensemaking ability.



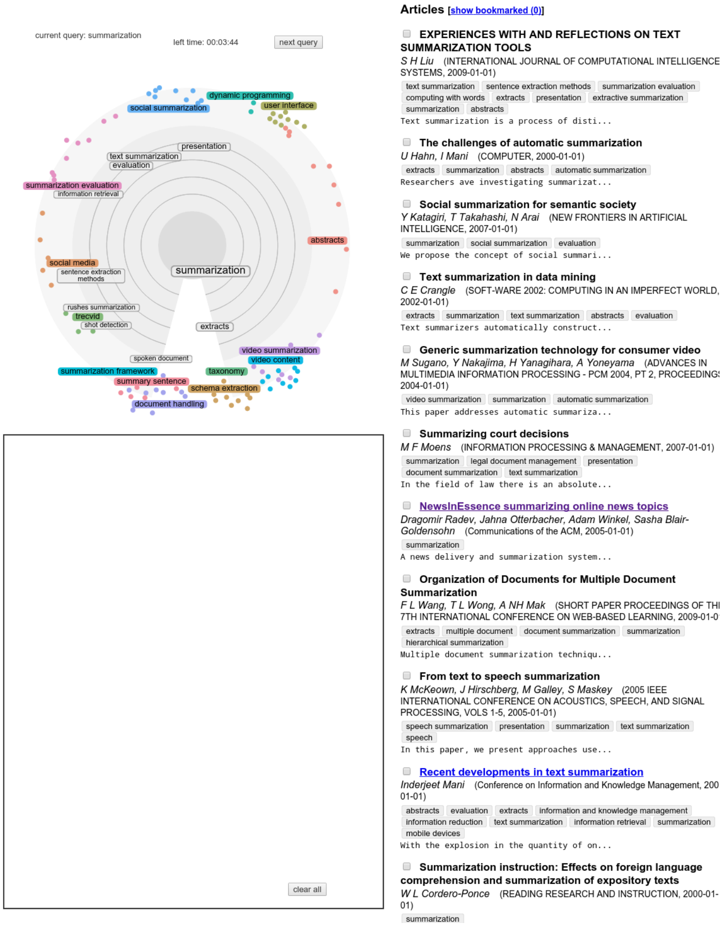
current query: summarization

left time: 00:03:44

Figure 6.1: the information panel replacing the query input box

All the important actions performed by users during the experiment will be logged and stored as in the experiment in Chapter 5. The actions are listed in Table 5.3.

The two systems with the information panel are shown in Figure 6.2 and Figure 6.3. Figure 6.2 shows the complete version of search system and Figure 6.3 shows the baseline version of the search system without the Intent Radar.



Articles [show bookmarked (0)]

☐ **EXPERIENCES WITH AND REFLECTIONS ON TEXT SUMMARIZATION TOOLS**
S H Liu (INTERNATIONAL JOURNAL OF COMPUTATIONAL INTELLIGENCE SYSTEMS, 2009-01-01)
text summarization | sentence extraction methods | summarization evaluation | computing with words | extracts | presentation | extractive summarization | summarization | abstracts
`Text summarization is a process of disti...`

☐ **The challenges of automatic summarization**
U Hahn, I Mani (COMPUTER, 2000-01-01)
extracts | summarization | abstracts | automatic summarization
`Researchers ave investigating summarizat...`

☐ **Social summarization for semantic society**
Y Katagiri, T Takahashi, N Arai (NEW FRONTIERS IN ARTIFICIAL INTELLIGENCE, 2007-01-01)
summarization | social summarization | evaluation
`We propose the concept of social summari...`

☐ **Text summarization in data mining**
C E Crangle (SOFT-WARE 2002: COMPUTING IN AN IMPERFECT WORLD, 2002-01-01)
extracts | summarization | text summarization | abstracts | evaluation
`Text summarizers automatically construct...`

☐ **Generic summarization technology for consumer video**
M Sugano, Y Nakajima, H Yanagihara, A Yoneyama (ADVANCES IN MULTIMEDIA INFORMATION PROCESSING - PCM 2004, PT 2, PROCEEDING
2004-01-01)
video summarization | summarization | automatic summarization
`This paper addresses automatic summariza...`

☐ **Summarizing court decisions**
M F Moens (INFORMATION PROCESSING & MANAGEMENT, 2007-01-01)
summarization | legal document management | presentation | document summarization | text summarization
`In the field of law there is an absolute...`

☐ **NewsInEssence summarizing online news topics**
Dragomir Radev, Jahna Otterbacher, Adam Winkel, Sasha Blair-Goldensohn (Communications of the ACM, 2005-01-01)
summarization
`A news delivery and summarization system...`

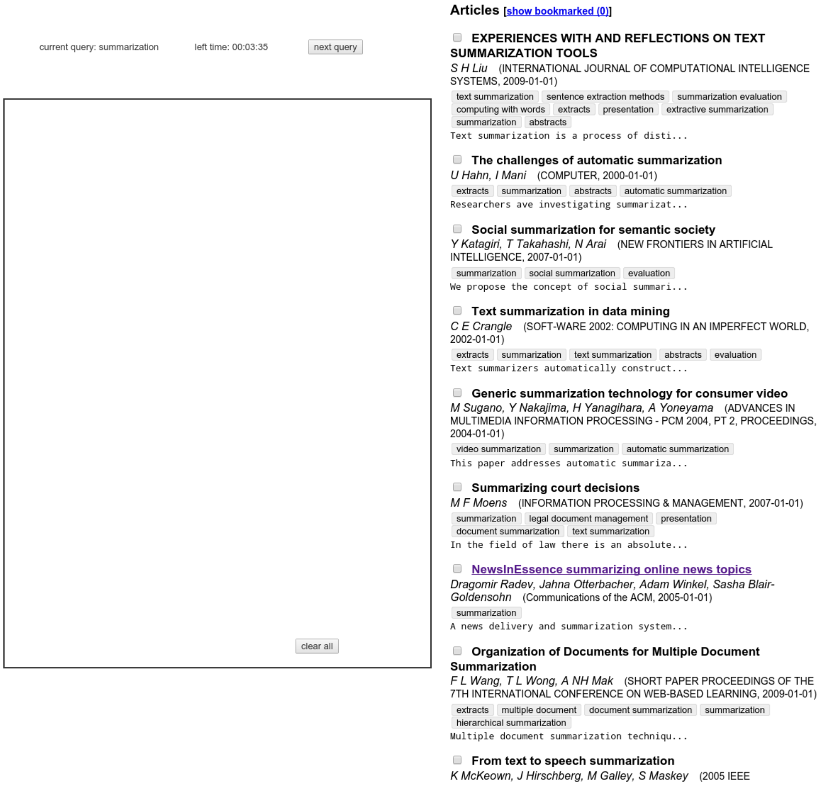
☐ **Organization of Documents for Multiple Document Summarization**
F L Wang, T L Wong, A NH Mak (SHORT PAPER PROCEEDINGS OF TH
7TH INTERNATIONAL CONFERENCE ON WEB-BASED LEARNING, 2009-01-0
extracts | multiple document | document summarization | summarization | hierarchical summarization
`Multiple document summarization techniqu...`

☐ **From text to speech summarization**
K McKeown, J Hirschberg, M Galley, S Maskey (2005 IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, VOLS 1-5, 2005-01-01)
speech summarization | presentation | summarization | text summarization | speech
`In this paper, we present approaches use...`

☐ **Recent developments in text summarization**
Inderjeet Mani (Conference on Information and Knowledge Management, 200
01-01)
abstracts | evaluation | extracts | information and knowledge management | information reduction | text summarization | information retrieval | summarization | mobile devices
`With the explosion in the quantity of on...`

☐ **Summarization instruction: Effects on foreign language comprehension and summarization of expository texts**
W L Cordero-Ponce (READING RESEARCH AND INSTRUCTION, 2000-01-01)
summarization

Figure 6.2: a complete version of the interactive exploratory search system



Articles [show bookmarked (0)]

□ **EXPERIENCES WITH AND REFLECTIONS ON TEXT SUMMARIZATION TOOLS**
S H Liu (INTERNATIONAL JOURNAL OF COMPUTATIONAL INTELLIGENCE SYSTEMS, 2009-01-01)
text summarization | sentence extraction methods | summarization evaluation | computing with words | extracts | presentation | extractive summarization | summarization | abstracts
`Text summarization is a process of disti...`

□ **The challenges of automatic summarization**
U Hahn, I Mani (COMPUTER, 2000-01-01)
extracts | summarization | abstracts | automatic summarization
`Researchers ave investigating summarizat...`

□ **Social summarization for semantic society**
Y Katagiri, T Takahashi, N Arai (NEW FRONTIERS IN ARTIFICIAL INTELLIGENCE, 2007-01-01)
summarization | social summarization | evaluation
`We propose the concept of social summari...`

□ **Text summarization in data mining**
C E Crangle (SOFT-WARE 2002: COMPUTING IN AN IMPERFECT WORLD, 2002-01-01)
extracts | summarization | text summarization | abstracts | evaluation
`Text summarizers automatically construct...`

□ **Generic summarization technology for consumer video**
M Sugano, Y Nakajima, H Yanagihara, A Yoneyama (ADVANCES IN MULTIMEDIA INFORMATION PROCESSING - PCM 2004, PT 2, PROCEEDINGS, 2004-01-01)
video summarization | summarization | automatic summarization
`This paper addresses automatic summariza...`

□ **Summarizing court decisions**
M F Moens (INFORMATION PROCESSING & MANAGEMENT, 2007-01-01)
summarization | legal document management | presentation | document summarization | text summarization
`In the field of law there is an absolute...`

□ **NewsInEssence summarizing online news topics**
Dragomir Radev, Jahna Otterbacher, Adam Winkel, Sasha Blair-Goldensohn (Communications of the ACM, 2005-01-01)
summarization
`A news delivery and summarization system...`

□ **Organization of Documents for Multiple Document Summarization**
F L Wang, T L Wong, A NH Mak (SHORT PAPER PROCEEDINGS OF THE 7TH INTERNATIONAL CONFERENCE ON WEB-BASED LEARNING, 2009-01-01)
extracts | multiple document | document summarization | summarization | hierarchical summarization
`Multiple document summarization techniqu...`

□ **From text to speech summarization**
K McKeown, J Hirschberg, M Galley, S Maskey (2005 IEEE

Figure 6.3: an baseline version of the interactive exploratory search system

6.2.3 Participants

Participants were recruited from the Department of Information and Computer Science, School of Science. During the participant recruitment, participants were required to fill in a questionnaire. in which basic personal information was collected, including name, email, gender, age, familiarity with the sciNet system, education level and English proficiency. After that, participants were required to evaluate their knowledge on the 24 selected topics discussed in Section 6.2.1, on a scale of one to five: level one means that participants are totally ignorant about this concept and know nothing about it; level two means that participants have basic impression this concept; level three means that participants have some modest knowledge of this concept; level four means that participants once had some courses about this concept before, otherwise on similar knowledge level; level five means that participants have done a lot work on this concept and are on an expert level. The topic selection for the participant follows the principle that the participant should not have zero knowledge on the topics and nor should they be experts on the topics. In order to follow that principle, only topics where they had some but not too much expertise were selected for them. Participants were not assigned with topics which they marked as level four or five because they would easily find out related concept even without the search system. Neither would they be assigned with topics marked as level one since it will be too difficult for them to perform. Ideally, participants would work on the topics that they are at the level two or three.

Ten participants were recruited in total and nine of them take part in the experiment. A pilot experiment was initially conducted with three participants in order to obtain a proper time setting of the experiment.[1] After that, six participants conducted the final experiment. Among the six participants, half are master-level students and half are doctoral students. The gender ratio between male and female is two to four. Five of them fall into the 25-32 age group and one is in the 33-40 age group. None of them is a native English speaker. Five of them have never heard of the Scinet system and one only heard about the system briefly. All of them are familiar with search engines and scientific search.

Right before the experiment, tutorial videos were shown to the participants in order to illustrate how to use the designed search system . The videos were shown to ensure that every participant had the same level of skills to utilize the systems.

6.2.4 Data

The dataset used in this experiment is the same dataset as used in paper [17]. This dataset contains over 50 million scientific articles from the Digital Libraries of the Association of Computing Machinery (ACM), the Institute of Electrical and Electronics Engineering (IEEE), the Web of Science prepared by Thomson Reuters and Springer. The arXiv documents are not used in the user experiment for two reasons. The first is the lack of sufficient expert knowledge to evaluate tasks suitable for arXiv. We only have expert knowledge in computer science but over 60% of the documents in arXiv are from the physics field and the articles in the field of computer science only occupy about 6%. The second reason is that we try to isolate the effect of the extracted keywords. The keywords for the arXiv articles are extracted

[1] The time limitation for each task was increased to five minutes (two minutes to collect information + three minutes to write answers) from three minutes (one minute to collect information + two minutes to write answers) based on the pilot experiment since we found out the three minutes setting was too short for the participant to conduct the task.

from the AKE algorithm and the quality of such keywords could affect the experiment in hard-to-control ways.

6.2.5 Evaluation and Measurement

In the evaluation process, we will first measure the relevance of the main concepts listed in the answers to the general topic and relevancies of the sub-concepts to their user-indicated main concept. The measurement is on a 0-10 scale. Rating 10 means that the main concept is fully relevant to the general topics or the sub-concept is fully relevant to the corresponding main concept. Rating 0 means that the main concept is not relevant to the general topic at all or the sub-concept is not relevant to the main concept at all. For each general topic, a rating was assigned to each main concept suggested by any user for that topic. For each main concept suggested within a topic, a rating was assigned to each sub-concept suggested by any user for that main concept. The ratings were thus given over the pooled set of answers rather than inspecting each participant.

After giving ratings for the listed main concepts and the sub-concepts, we will evaluate the answers for the general topic from two perspectives, breadth and depth. The breadth of the answer is represented by relevance of the user's listed main concepts (MC) to its general topic (GT). The breadth for the *ith* general topic of the *mth* participant is calculated as follows:

$$Breadth(GT_i^m) = \sum_j (MC_{ij}^m)$$

in which GT_i^m means the *ith* general topic of the *mth* participant and MC_{ij}^m is the rating for the *jth* main concept listed under the general topic of the *mth* participant. This formula means that we represent the breadth as the summation of expert-given weights for all the main concepts listed by the user under the general topic.

The depth of the answer is represented by the depth of the listed main concepts. The depth is evaluated from two perspectives: precision and recall. They are calculated from the relevance rating of the sub-concepts as follows:

$$Precision(MC_i^m) = \sum_j (LSub_{ij}^m)/(\#(LSub_i^m) \times 10)$$

$$Recall(MC_i^m) = \sum_j (LSub_{ij}^m)/\sum_j (LSub_{ij}^{all})$$

in which MC_i^m is the *ith* main concept listed by the *mth* participant (the index of the general topic is omitted for clarity), $LSub_{ij}^{all}$ is the rating of the

jth sub-concept listed overall by all participants for main concept i.[2] and $\#(LSub_i^m)$ is the total number of the listed sub-concepts under the ith main concept of the mth participant.

After calculating the precision and recall value for each of the main concepts listed under the general topic, the average precision and recall value will be both used as the indicator of the depth measurement of the topic.

Each participant works on eight general topics in total, including four tasks for the full version of the system and four tasks for the baseline version of the system. The breadth and depth analysis will be conducted on all the tasks. Then the average performance of the four tasks on the same system, which is the average breadth and depth, will be calculated and used to demonstrate which system helps the participant achieve better performance. The same evaluation process will go through all the six participants. Based on the performance over all the participants, we can then judge whether the interactive search system helps to assist the sensemaking process. Particularly, since there are few participants, all the values will be listed instead of taking an average over participants.

We make the hypothesis that the IESS supports the sensemaking process. Under this hypothesis, we expect that the participants will make more sense and understand the general topics better on the full system.

6.2.6 Results and Conclusion

The evaluation of the experiment result is based on three measurements, including breadth, precision and recall as detailed in the previous subsection. Table 6.2 lists the average breadth of the answers from both of the systems. From the table, we can see that four out of six participants achieved better performance on the full system in term of breadth.

Table 6.2: breadth of the answers from Participant 1 to 6 (P1 to P6)

	p 1	p 2	p 3	p 4	p 5	p 6
Full system	22.5	22.5	20.25	20.75	24.5	17.25
Baseline system	20	22.25	22	23	13	16.5

Table 6.3 shows the average precision of the answers from both of the systems. From the table, we can see that only two out of six participants achieved better precision on the full system and the baseline system are more likely to help the participants to obtain more precise answers.

[2] Note that the set of sub-concepts listed overall by all participants is larger than the set listed by any individual participant.

Table 6.3: precision of the answers from Participant 1 to 6 (P1 to P6)

	p 1	p 2	p 3	p 4	p 5	p 6
Full system	0.62	0.50	0.52	0.38	0.65	0.41
Baseline system	0.26	0.64	0.66	0.4	0.53	0.65

Table 6.4 presents the average recall of the answers from both of the systems. From the table, we can see that four out of six people achieved better performance on the full system in term of recall.

Table 6.4: recall of the answers from Participant 1 to 6 (P1 to P6)

	p 1	p 2	p 3	p 4	p 5	p 6
Full system	0.72	0.70	0.92	0.52	0.64	0.62
Baseline system	0.59	0.66	0.86	0.40	0.82	0.90

Precision of the answers is the reflection of the understanding of the general topic and measures the accuracy of the description for the user's listed main concepts. Thus they should achieve higher precision on the full system. Breadth and recall tell how much of the important concepts the user has noticed for the main concepts. Moreover, as the IESS is initially designed for supporting exploratory search and broadening the search space, obviously, we can expect that participants will achieve higher breadth and recall on the full system.

The experiment result shows that participants obtain lower precision, higher breadth and higher recall on the full system compared to the baseline system. Based on this observation, we can conclude that it is uncertain whether the IESS can help the users to increase the understanding of a topic due to the lower precision. However, the higher breadth and recall indicate that using IESS has helped the users discover more important concepts compared with using the baseline system.

Although the experiment results are not fully consistent with our expectations, it is also worth mentioning that the IESS received positive feedback from the participants during the feedback session after the experiment. Most of the participants said that the IESS had helped them to answers the questions and it was more difficult to conduct the tasks without the Intent Radar. This implies that the IESS had made its contribution to the sensemaking process.

Chapter 7

Overall Conclusion and Future Work

In this thesis, two main contributions have been made concerning interactive exploratory search systems.

The first contribution is that an interactive exploratory search system (IESS) has been developed for the arXiv database and a small scale user experiment has been conducted based on the arXiv dataset. The experiment result has shown that the IESS has the functionality to support interactive exploratory search. However, as discussed in Chapter 5, the system still needs to be improved. In the future, it is of crucial importance to improve the quality of the extracted keywords. Currently, KP-Miner is used as the keyword extraction tool. However, it is unsatisfactory as shown in practice. For example, many common words are recognized as keywords. Another AKE tool must be found to improve the keyword quality.

The second contribution is that a user experiment has been designed and conducted to investigate the hypothesis that the interactive exploratory search system could support sensemaking due to the keyword relationships shown on the Intent Radar. The user experiment result preliminarily suggests that the IESS has improved the breadth and recall at some cost to precision. Based on the feedback of the participants, the Intent Radar was useful to help the participants to answer the questions. It is also worth mentioning that all the participants' operations on the systems during the experiment are logged. The recorded operations can be found in the Table 5.3. In the future, these log files can be analysed to detect the participants' behaviour patterns, allowing more detailed investigation of how the system could support the sensemaking process.

Bibliography

[1] Abraham, A., Petre, M., and Sharp, H. Information seeking: Sensemaking and interactions. *City* (2008).

[2] Agosti, M. Future in information retrieval evaluation. *Evaluation Methodologies in Information Retrieval*, 102.

[3] Auer, P. Using confidence bounds for exploitation-exploration tradeoffs. *The Journal of Machine Learning Research 3* (2003), 397–422.

[4] Azcarraga, A., Liu, M. D., and Setiono, R. Keyword extraction using backpropagation neural networks and rule extraction. In *Neural Networks (IJCNN), The 2012 International Joint Conference on* (2012), pp. 1–7.

[5] Bohne, T., and Borghoff, U. M. Data fusion: Boosting performance in keyword extraction. In *Engineering of Computer Based Systems (ECBS), 2013 20th IEEE International Conference and Workshops on the* (2013), IEEE, pp. 166–173.

[6] El-Beltagy, S. R., and Rafea, A. Kp-miner: A keyphrase extraction system for english and arabic documents. *Information Systems 34*, 1 (2009), 132–144.

[7] Fisher, K., Counts, S., and Kittur, A. Distributed sensemaking: improving sensemaking by leveraging the efforts of previous users. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2012), ACM, pp. 247–256.

[8] Kang, Y.-B., Haghighi, P. D., and Burstein, F. Cfinder: An intelligent key concept finder from text for ontology development. *Expert Systems with Applications* (2014).

[9] Kim, S., Medelyan, O., Kan, M.-Y., and Baldwin, T. Automatic keyphrase extraction from scientific articles. *Language Resources and Evaluation 47*, 3 (2013), 723–742.

[10] Klein, G., Moon, B. M., and Hoffman, R. R. Making sense of sensemaking 1: Alternative perspectives. *IEEE intelligent systems 21*, 4 (2006), 70–73.

[11] Lopez, P., and Romary, L. Humb: Automatic key term extraction from scientific articles in grobid. In *Proceedings of the 5th international workshop on semantic evaluation* (2010), Association for Computational Linguistics, pp. 248–251.

[12] Marchionini, G. Exploratory search: from finding to understanding. *Communications of the ACM 49*, 4 (2006), 41–46.

[13] Perry, J., Janneck, C. D., Umoja, C., and Pottenger, W. M. *Supporting Cognitive Models of Sensemaking in Analytics Systems* (2009).

[14] Pirolli, P., and Russell, D. M. Introduction to this special issue on sensemaking. *Human-Computer Interaction 26*, 1 (2011), 1–8.

[15] Romero, M., Moreo, A., Castro, J. L., and Zurita, J. M. Using wikipedia concepts and frequency in language to extract key terms from support documents. *Expert Systems with Applications 39*, 18 (2012), 13480–13491.

[16] Rose, S., Engel, D., Cramer, N., and Cowley, W. Automatic keyword extraction from individual documents. *Text Mining* (2010), 1–20.

[17] Ruotsalo, T., Peltonen, J., Eugster, M., Glowacka, D., Konyushkova, K., Athukorala, K., Kosunen, I., Reijonen, A., Myllymaki, P., Jacucci, G., and Kaski, S. Directing exploratory search with interactive intent modeling. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management* (2013), ACM, pp. 1759–1764.

[18] Sarkar, K. A hybrid approach to extract keyphrases from medical documents. *arXiv preprint arXiv:1303.1441* (2013).

[19] Turney, P. D. Learning algorithms for keyphrase extraction. *Information Retrieval 2*, 4 (2000), 303–336.

[20] Venna, J., Peltonen, J., Nybo, K., Aidos, H., and Kaski, S. Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *The Journal of Machine Learning Research 11* (2010), 451–490.

[21] Vidal, M., Menezes, G. V., Berlt, K., Moura, E. S. D., Okada, K., Ziviani, N., Fernandes, D., and Cristo, M. Selecting keywords to represent web pages using wikipedia information. In *18th Brazilian Symposium on Multimedia and the Web, WebMedia 2012* (2012), ACM, pp. 375–382.

[22] Wartena, C., Brussee, R., and Slakhorst, W. Keyword extraction using word co-occurrence. In *7th International Workshop on Text-Based Information Retrieval, TIR 2010 - In Conjunction with DEXA 2010* (2010).

[23] Wildemuth, B. M., and Freund, L. Assigning search tasks designed to elicit exploratory search behaviors. In *Proceedings of the Symposium on Human-Computer Interaction and Information Retrieval* (2012), HCIR '12, ACM, pp. 4:1–4:10.

[24] Wilson, M. J., and Wilson, M. L. A comparison of techniques for measuring sensemaking and learning within participant-generated summaries. *Journal of the American Society for Information Science and Technology 64*, 2 (2013), 291–306.

[25] Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C., and Nevill-Manning, C. G. Kea: Practical automatic keyphrase extraction. In *Proceedings of the fourth ACM Conference on Digital Libraries* (1999), ACM, pp. 254–255.

[26] You, W., Fontaine, D., and Barthes, J.-P. An automatic keyphrase extraction system for scientific documents. *Knowledge and Information Systems 34*, 3 (2013), 691–724.