1 Project Details

1.a) Project Title. Content-based Literature Search using Knowledge and Structure

1.b) Project Acronym. CLiKS

1.c) Principal Investigator. Prof dr. Maarten de Rijke
Intelligent Systems Lab Amsterdam (ISLA)
University of Amsterdam
E-mail: mdr@science.uva.nl
URL: http://ilps.science.uva.nl/

1.d) Renewed application. This is *not* a renewed application.

2 Summary and Dutch language abstract

2.a) Summary. The project aims to develop new retrieval models and algorithms for searching and browsing in scientific literature. Two important recent developments in the scientific literature production process form the concrete motivation for this project: (1) semantically rich document structuring standards, and (2) increasingly rich keyword annotations that capture domain knowledge. The driving question underlying this proposal is: How can we use these to improve access to scientific literature? To address this question we propose to use rich probabilistic retrieval models that allow us to capture the relation between document content, document structure, and document-level annotations.

To provide focused access, we aim to return semantically defined XML elements, with query models informed by available domain knowledge. We will contrast generative language modeling based approaches with approaches based on discriminative models that may allow for better optimization and estimation methods.

Evaluation is done using standard benchmarks provided by INEX and TREC. On top of that a richly marked up and annotated corpus provided by a leading scientific publisher will be used for a system-centered comparison and for a user study on the benefits of semantically oriented markup vs layout-oriented markup.

2.b) Abstract for laymen (in Dutch). Het project heeft als doel het ontwikkelen van nieuwe zoekmodellen en -algoritmes bestemd voor het zoeken in en browsen door wetenschappelijke literatuur. Twee belangrijke recente ontwikkelingen in het productieproces van wetenschappelijke literatuur vormen de concrete aanleiding voor dit project: (1) semantisch rijke documentstructuurstandaarden en (2) steeds betekenisvollere trefwoorden die domeinkennis aan documenten toevoegen. De onderliggende vraag van dit voorstel is hoe we deze ontwikkelingen kunnen toepassen om toegang tot wetenschappelijke literatuur te verbeteren. Om deze vraag te kunnen beantwoorden maken we gebruik van probabilistische retrievalmodellen welke ons de mogelijkheid geven de relatie tussen documentinhoud, documentstructuur en annotaties op documentniveau te combineren.

Om precieze, gerichte toegang te geven tot informatie in documenten proberen we semantische XML elementen als zoekresultaat te geven, door gebruik te maken van query modellen die verrijkt zijn met domeinkennis. We vergelijken methodes gebaseerd op generatieve language models met discriminatieve language models, waarbij deze laatste beter geschikt zouden kunnen zijn voor optimalisatie- en schattingmethodes.

De standaardbenchmarks van INEX en TREC worden gebruikt voor evaluatie. Daarnaast wordt een rijk geannoteerde en gestructureerde documentcollectie gebruikt, welke beschikbaar gesteld wordt door een belangrijke wetenschappeljke uitgever. Deze collectie wordt gebruikt voor systeem-gecentreerde vergelijkingen en voor een gebruikersstudie naar de voordelen van semantische structuur versus layout structuur.

3 Classification

The project falls within the discipline of *Computer Science*. Relevant *Nationale Onderzoeksagenda Informatie- en Communicatietechnologie 2005-2010 (NOAG-ict)* research themes:

3.2 Data Explosie. ICT-disciplines: Algorithms and Computation Theory; Hypermedia, Hypertext and Web; Information Retrieval; Knowledge Discovery in Data.

3.3 Digitale Beleving ICT-disciplines: Computer-Human Interaction; Information Retrieval.

4 Composition of the Research Team

Name	Title	Role	Expertise	Affiliation
de Rijke, M.	Prof.dr.	Project leader	information storage and retrieval	ISLA, U. Amsterdam
Kircz, J.G.	Dr.	Project leader	scientific publishing electronic publishing	KRA Publishing Research Inst. for Media & Inf. Mgt, Hogeschool van Amsterdam
Sieverts, E.G.	Dr.	Collaborator	retrieval systems	Utrecht U. Library
Goris, G.	Drs.	Collaborator	digital libraries	Erasmus U. Rotterdam Library
Schrauwen, R.	Dr.	Collaborator	electronic publishing	Central App. Mgt, Elsevier
de Waard, A.	Drs.	Collaborator	electronic publishing	Adv. Technology Grp, Elsevier
de Belder, K.	Drs.	Advisor	electronic publishing	Leiden U. Library
Fuhr, N.	Prof.dr.	Advisor	digital libraries, XML retrieval	Universität Duisburg
NN		PhD student	scientific information retrieval	To be funded by NWO

The project seeks to employ a PhD student. The project will be directed by De Rijke, who will act as formal promotor for the PhD student, and who will oversee the retrieval aspects of the research to be carried out. Kircz will oversee the digital library aspects of the project, and will act as co-promotor.

The collaborators from the university libraries will oversee and inform the user studies we foresee, and they will carry out comparisons between results produced by the project's search aids and existing search aids. The collaborators from Elsevier provide input and feedback from the digital library production and access points of view.

The advisors listed will not be involved with the project on an ongoing basis, but they have agreed to be available for high-level consultation, general strategic advice, and for evaluation of the project's outcomes.

5 Research School

The research will be embedded in the Dutch Research School for Information and Knowledge Systems (SIKS).

6 Description of the Proposed Research and Application Perspective

6.a) Description of the Proposed Research

a. Research Questions and Desired Results

The problem we address is access to scientific publications. Scientific researchers have difficulties keeping current with new research findings that continue to grow at an exponential rate [35]. In the setting of a digital library information access is usually a mixture of two tasks: *searching* and *browsing* [23, 32, 57]. Searching is associated with locating documents that meet the information need a user has expressed in terms of a query. When browsing, users glimpse a field, select or sample an informational object, and examine or abandon it [9]. By browsing, the user is able to build up a conceptual map of the information need at hand, one that can help him to articulate specific queries or to broaden the search. In a digital library environment, browsing is often facilitated through a form of controlled vocabulary.

Most information systems developed to facilitate scientific discovery (e.g., [6, 48, 55]), are based on traditional search and browsing approaches that build on terms, keywords and/or subject indexing. Increasingly, users adopt a "locate-and-read" strategy instead of the more traditional "read-and-locate" typical of a paper environment. To support this type of access, search engines need to move from simply returning

relevant documents to providing "go-read-here" functionality, where readers are given focused access—searching and browsing—to highly relevant sub-parts of documents in context.

Against this background, two developments in today's scientific literature production process are particularly relevant [3, 30]. First, increasingly, today's scientific documents are equipped with explicit coding for *document structure*. Obviously, all text has structure, but the degree of "structured-ness' varies between documents and document collections. In state of the art scientific literature production processes, text is explicitly marked up in XML, the eXtensible Markup Language, both to record layout cues and to provide some amount of semantics—through subject terms, metadata, bibliographic links, and, potentially, "typed" sections. Second, as a consequence of the renewed interest in controlled vocabularies, thesauri, and other ways of explicitly modeling domain knowledge (such as ontologies), scientific articles are being marked up with increasingly rich *keywords*. The "richness" may reside both at the document end (with more and/or hierarchically organized keywords being assigned) or, more likely, at the backend, where keywords are part of elaborately organized structures; the medical subject headings (MeSH) controlled subject vocabulary [37] provides a well-known example in the medical domain.

The overall *aim* of the **CLiKS** project is to develop flexible and robust retrieval models for scientific literature search that allow for arbitrary document and text features to be incorporated as evidence in the search and discovery process. In particular, we want to make use of document structure and keyword annotations as well as textual features based on occurrences of terms and keywords in particular elements. To address our overall aim, we will pursue the following *research questions*:

RQ1 How can we model retrieval at the element-level, as the element provides the context and semantics of the words it contains? To what extent are explicit XML tags—either content-oriented or layout-oriented—roads signs in this quest?

RQ2 How can we model retrieval of documents annotated with keywords taken from rich domain models, taking account both of concepts captured by the model and of relations between those concepts?

RQ3 How can retrieval models that capture rich document-level keyword annotations be combined with element-level evidence?

RQ4 How can retrieval of scientific literature (enriched with keyword annotations and document structure) be evaluated in context?

The *desired results* of the project come in three kinds:

- *Models, algorithms, and estimation methods* (for searching and browsing semi-structured, richly annotated full length articles),
- *Methodology* (studying the context of scientific literature search and browsing, test sets, experimental environment),
- *Software* (demonstrator system for scientific literature search and browsing, released under GPL)

Thus, while our main focus will be *system-centered* aspects of scientific literature retrieval, with both theoretical and experimental components, *user-centered* aspects will also be addressed, in context.

b. Approach and Methodology

In this section we present our overall approach to the problem of content-based access to scientific literature that exploits the knowledge annotations and semantic structure that accompany today's scientific literature. Following [56], we view theoretical, experimental, and applied aspects of academic research as inextricably interlinked, where each builds on, and feeds into the other. To facilitate the desired interaction, the project will aim to a have baseline prototype from early on in the project; this will be possible as baseline versions of the required components are available at the host institute; see Section 8 below for details.

The planned research on system-centered aspects of the project will contrast two approaches towards incorporating arbitrary document and text features as evidence for computing relevance. More generally, an important desirable feature in modern information retrieval (IR) is its ability to incorporate a broad range of features derived from collections and documents. There may also be query independent features that influence the relevance of a document (or document fragment) [41].

In the so-called generative language modeling approach to IR, a model is trained on each document in the collection to be searched. The ranking of a document is given by the probability of generating the query from the document's language model. Language models have been quite successful in several IR tasks and their performance has been shown to be on a par with the state-of-the-art. The ISLA team has

considerable experience with retrieval based on language models for scenarios or tasks that require some form of combination of evidence, either using mixture models or priors, with weights learned automatically using the EM algorithm or by empirical means. Examples include scientific literature search, web search, enterprise search, XML retrieval, ad hoc retrieval, blog retrieval, etc.

Discriminative models have been widely used for pattern classification [16]. In recent years, they have been introduced into a variety of language technology tasks [14], and they have also been applied to IR tasks [20, 41, 60]. Their main advantages are avoiding unrealistic modeling assumptions (mainly concerning term independence [20]), expressiveness, being able to directly model classification problems (i.e., into relevant vs. non-relevant) and to directly optimize for performance measures, as well as allowing for the integration of arbitrary features that influence relevance.

To address RQ1 (Modeling structure) we will start from our existing work on language modeling for XML element retrieval and we will contrast this with a discriminative model for XML element retrieval (to be created within the project). For evaluation purposes we will make use of standard benchmarks. In addition we aim to contrast, within a retrieval setting, document structure (and markup) purely based on layout markup and elements vs. semantically defined markup and elements (such as "introduction"). To perform evaluation we need to create a dedicated test set; for this purpose Elsevier, one of the world's largest scientific publishers, has made available a document collection marked up using Elsevier's DTD 5 family (see Section 8 for details). The project will create a test set from this document collection, using topic development guidelines made available by the INEX initiative [25].

To address RQ2 (Modeling knowledge) we will again contrast a generative and discriminative approach. We start with the approach to integrating (purely) conceptual knowledge in generative LM setting as described by [4] and turn this into a discriminative approach using the aproach sketched by Gao et al. [20]. Our existing LM-based approach is based on two-stage query models, in which concepts act as an intermediate between query terms and documents. Next, we will turn to integrating not just the concepts but also the relations between concepts into the modeling, allowing us to expand the set of concepts and, indirectly, the set of terms related with a given query. To model "related concepts" in a probabilistically sound manner, we will consider various options, both graph-based and based on standard information theory measures. For evaluation purposes, we will make use of the TREC genomics test collections; see Section 8.

For RQ3 (Modeling knowledge and structure), the main challenge is to combine the features considered for RQ1 and RQ2, i.e., document level features (such as keyword annotations), element level features (such as specific types of markup), and text features. We will pursue several ways of inducing element level keyword annotations from document level ones, starting with standard measures from information theory (e.g., log-likelihood or KL-divergence), viewing the original documents (whose annotations need to be "pushed down to the element level") as a reliable source of language usage typically associated with the keyword(s) at hand. We will subsequently follow a dual approach to (element) retrieval, again, both generative and discriminative. For evaluation purposes we will re-use the test set created for RQ1.

While TREC-style system-centered evaluations are extremely useful for system development, they do not inform us about the ways in which real users appreciate the ranking features offered by the retrieval models that we will develop in the course of this project [26, 47]. For our final research question, RQ4 (user-centered evaluation), we change track and conduct a user-based evaluation instead of a system-centered evaluation. A prototype based on the best performing model to come out of our investigations of RQ3 will be made available to small groups of users at the participating university libraries. We employ participatory design [40, 50] as a sanity check, so as to ensure that the demonstrator is not orthogonal to the needs of our user populations. Specifically, in our experimental setup we will adopt the framework of the INEX interactive track [33], whose simulated task setup we have found very instructive in earlier work [52, 54].

c. Significance and Urgency

Scientific literature search using content, knowledge, and document structure is extremely timely as a research topic since it brings together three ongoing developments in the field of information retrieval: renewed interest in new probabilistic retrieval models, the move to semantically-oriented document structure in the scientific literature production process, and the use of ever richer domain knowledge to improve retrieval effectiveness.

Against this general background, the main points demonstrating the significance and urgency of the proposed research are: (1) the continued interest by the research community as well as applied partners in scientific literature search, as witnessed by the TREC genomics track, the INEX book search task, and the CLEF 2008 ad hoc track (see below); (2) a recognized need for retrieval research that complements

system-centered work with human-centered evaluation that translates findings of human-centered research into models and algorithms for IR [26, 47]; (3) the hosting institute's unique position to address these challenges, because we have people familiar with literature search, XML retrieval, biomedical retrieval, probabilistic models for IR, and human-centered research; (4) our findings will be generalizable to other types of information seeking and IR research because we build on existing theoretical frameworks and methodology.

Specific points to attest to the significance of our proposal for the area of scientific literature search are the fact that we will be working with data provided by one of the largest scientific publishers and that we take our theoretical findings out of the controlled laboratory setting to real-world settings in context, through the collaboration of, and interaction with, 3 large academic libraries.

Finally, in a recent overview [34] provide a taxonomy of XML retrieval use cases; it emerges that most research so far has concerned *layout-oriented document types* and *process-oriented document types*, with relatively little work on *content-oriented document types* of the kind that **CLiKS** will be examining.

d. Related Work

Related work comes in several kinds: scientific literature search, XML element retrieval, biomedical search, user-based evaluation, and new types of retrieval models. We discuss each in turn.

Scientific literature search. The study of the information seeking behavior of scientists can be traced back to the 1940s [18, 19, 22, 46]. One of the constant themes has been the metadata (and controlled vocabulary) vs. full-content indexing opposition. We explicitly opt for *both*, and we strongly believe in building on insights from either camp in our efforts to improve the quality of information access. We believe that this combined strategy—a distinguishing feature of our proposal—is the way forward.

In comparing **CLiKS** with related work it is important to distinguish between research into access to certified content, non-certified content, and mixtures of the two. Citeseer [13] and Libra [42] provide access to non-certified scientific content crawled from the web. Google Scholar [21] offers access to both certified and non-certified content. **CLiKS** deals with certified content only. There is a sizeable number of ongoing initiatives aimed at providing access to certified scientific literature. Important examples include the Web of Science [58], Scopus [49], and ScienceDirect [48]. To the best of our knowledge none of these use a combination of content, metadata, and element-level document structure in the way that **CLiKS** proposes.

Reflecting the ongoing research interest in scientific literature search, later this year, CLEF, the Cross-Language Evaluation Forum, will feature a new main task for monolingual and cross-language search on library catalogue records; see `http://www.clef-campaign.org/`. The task is organised in collaboration with The European Library, but will most likely not feature full length articles, as a result of which **CLiKS** does not plan to take part.

XML element retrieval. With the arrival of INEX, the INitiative for the Evaluation of XML Retrieval, the study of retrieval of semi-structured documents, of sub-document units and of XML elements received a big boost. Held annually since 2002, INEX has spawned a great detail of interest in models, algorithms, and evaluation methodology for XML retrieval.

The UvA team was one of the first teams to use language modeling-based approaches to XML element retrieval at INEX [27, 28, 53]. Subsequently, various teams experimented with extensions of this basic language model. Graphical models have also been deployed at INEX (see e.g., [45] or Ogilvie and Callan [43] who performs smoothing using a weighted mixture of a background model, a document model, a parent model, and a mixture of the children models. A systematic analysis of discriminative modeling for XML element retrieval has not been performed yet, to the best of our knowledge.

Biomedical search. In recent years, search in biomedical literature has received an enormous boost. In 2003, reflecting the importance of this type of search, TREC launched a 5-year Genomics track in which increasingly challenging search tasks have been addressed; we refer to the track home page [59] for overviews and pointers to the work carried out there. One of the leading and recurring themes at the track has been the integration of rich domain knowledge in the retrieval process. Some participants used a largely heuristics based approach [62] while others went for a more principled language modeling-based approach [36].

User-based evaluation. User models for IR explicitly or implicitly model user interest, tasks, and goals, with interests being the most frequently modeled characteristic[17, 51]. Features explored so far have mainly been search specific, based on user queries, result clicks, or other page visits. Resulting user models are applied to personalize search results, for example through query expansion, relevance feedback, or result reranking. User models have mostly been static, focusing either on very short interaction cycles, or on modeling relatively static interests. Problems in combining long-term interests and short-term objectives have been explored but have not yet shown satisfying results [51, 61].

A closely related area of research is adaptive hypermedia (AH) [10, 15] which aims at developing interactive systems that automatically adapt content, navigation support, or presentation to users. Initially, research in User Modeling (UM) was driven by the requirements of a specific application and a wide range of user characteristics was explored in different contexts [31] (RQ4). Generic user modeling systems are evolving with the goal of providing reusable UM services that can be integrated with a number of applications. A major challenge remains to develop generalizable theories on which user characteristics can be modeled based on what features and what methods. While addressing this challenge of generalizability is beyond the scope of this proposal, our user study (RQ4) will be informed by ongoing developments [11, 24, 29, 44].

New types of retrieval models. Over the past few years, and driven by the need for flexible and robust retrieval models that are capable of incorporating a broad range of types of evidence, conferences such as SIGIR and CIKM have regularly featured publications on either very rich language modeling-based approaches to the document retrieval problem (with mixture models and priors) or approaches based on discriminative models. Examples have been given in 6.a.b.

e. Embedding

The proposed research continues the research on scientific literature search [1, 12], XML element retrieval [2, 5, 52, 54], biomedical literature search [4], and query and retrieval modeling [7, 8, 36] currently conducted at the host institute. As our approaches have been shown to achieve competitive results, a natural extension is to address limitations that prevent the technology from taking advantage of the features offered by today's scientific article production process. With experience in information retrieval in general and domain-specific retrieval in particular as well as user modeling and running large scale demonstrators (e.g., MoodViews.com, VerkiezingsKijker.nl), our team is in a unique position to address the challenges posed by scientific literature search.

The ILPS group within the Intelligent Systems Lab Amsterdam (ISLA) at the University of Amsterdam is one of the largest academic research groups in IR in Europe; it is a regular participant in, and co-organizer of, worldwide evaluation efforts such as TREC, CLEF, INEX, and NTCIR. ILPS co-organized SIGIR 2007 in Amsterdam, the world's leading IR conference.

6.b) Application Perspective

Results of this research are expected to be applicable to scientific literature search for a wide range of scientific disciplines, as long as semantically informed document structure and rich keyword annotations are available—given recent developments in the scientific literature production process this already holds true for many disciplines, and will soon hold true for many more.

We explicitly include the development of demonstrators in order to ensure that our research maintains a healthy balance between theory, experiment, and application. The first applications are expected during the project's lifetime. Developed software will be made available for public use. Knowledge transfer will happen through publications, both scientific and aimed at a general audience.

The extensive Elsevier DTD according to which the corpus made available by Elsevier is structured entails two strains. First, the XML tags are introduced to enable rendering the document on a great variety of output media. Second, the XML tags denote the information of relevant elements and structure. From a producer's point of view it is important to gain a better understanding as to which (types of) XML tags can be used for which usage. In other words, not all XML tags will add to a better retrieval of relevant information, because they are mainly introduced for lay-out purposes. Hence, the variation in value of the various kinds of tags is a natural outcome of our investigation. This information is an important ingredient for improving and adjusting the current XML DTD.

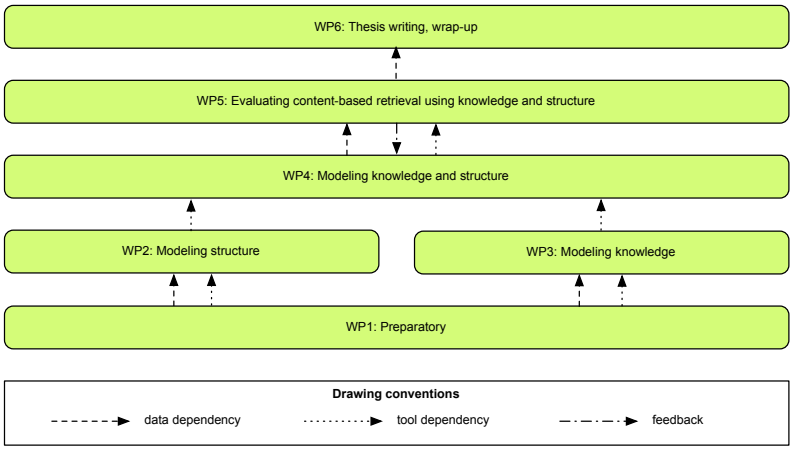

Figure 1: Main stages of the project.

Table 1: Work package overview and planning

WP	RQ	WP title	Start/End Month	Related milestones
WP1		Preparatory	1/6	I
WP2	RQ1	Modeling structure	7/15	II
WP3	RQ2	Modeling knowledge	16/24	III
WP4	RQ3	Modeling knowledge and structure	25/33	IV
WP5	RQ4	Evaluating content-based retrieval using knowledge and structure	34/42	IV, V
WP6		Thesis writing, wrap up	43/48	V

7 Project Planning

a. Project Structure

The project will run for 4 years starting in the fall of 2008. The project is structured into 6 work packages (WPs): a preparatory work package, 4 WPs corresponding to the 4 research challenges RQ1–RQ4, and the final thesis writing. Fig. 1 shows the project's organization and Table 1 provides an overview of its work packages and their planning.

The project will be developed in five R&D phases, each leading to a milestones and targeting the completion of (the implementation of) a specific work package; see Table 1. In the start-up phase the emphasis will be on platform development and establishing the baseline retrieval infrastructure and indexes; this will build on existing Lemur-based tools from UvA. **Milestone I** (month 6) is reached when the first implementation (with indexes for element and document level indexes for INEX, TREC genomics, and the Elsevier corpus) is available. **Milestone II** (month 15) will incorporate the modules developed in WP2; the focus of the milestone will be on *document structure*. During the next phase, leading to **Milestone III** (at month 24), we integrate keywords in the document-level retrieval approaches made available as part of Milestone 1. In the next phase, leading to **Milestone IV** (at month 33) the integrated version of the Milestone II and III systems will be delivered. During the final consolidation phase, month 34–48), with a **fifth and final milestone** at month 48, an evaluation of the Milestone IV system will be conducted; we perform a simulated task experiment in cooperation with three university libraries.

In each new R&D phase, the quality of the retrieval platform will be improved and extended; where needed, it will be made more robust or efficient.

b. Description of the Work Packages.

Below we describe each WP, indicating its goals and desired output, the dependencies on/from other WPs, the resources used by the WP and the possible risks and remedies.

WP1 Preparatory The objective of this work package is to setup a basic document/element retrieval framework based on Lemur and an evaluation infrastructure, together with a UI so that the retrieval en-

vironment can also be used for demonstration purposes. The system will form the point of departure for WP2, WP3 and WP4.

- *Output*: Literature survey on language modeling-based retrieval; baseline language modeling-based retrieval setup
- *Evaluation*: Reproduce earlier UvA results (based on a legacy framework)
- *Uses*: Lemur; TREC and INEX evaluation scripts; interface from XMLFind/Wikiii [52]
- *Risk*: Unable to reproduce old UvA results. *Remedy*: Invest in optimization and/or settle for median scoring system.

WP2 Modeling structure. Define XML element retrieval models based on a discriminative approach. For optimization and training purposes, compare different methods with simple sweeps of the parameter space. Start from feature-based models (in the style of Gao et al. [20] or Metzler and Croft [38]) and consider more involved models based on, e.g., [60].

- *Output*: Discriminative element retrieval engine; test set.
- *Evaluation*: Use earlier INEX test sets; develop a new test set based on the corpus made available by Elsevier (Section 8) for comparing layout-based vs content-based markup.
- *Dependencies*: Builds on WP1.
- *Uses*: Optimization methods described in [20, 39, 41, 60]. INEX topic development and result assessment guidelines.
- *Risk*: Discriminative models outperformed by more traditional language modeling-based ones. *Remedy*: Failure analysis; start like [20], i.e., close LM-based approach and slowly generalize from there. *Risk*: Test set creation too time-consuming. *Remedy*: involve undergraduate students for topic creation and assessment.

WP3 Modeling knowledge. In this WP we work at the document level, and first extend our language modeling approach so that related concepts (and, optionally, terms related to the expansion concepts) can be brought to bear on the retrieval process. Build a discriminative counterpart, and compare.

- *Output*: Models for expanding queries with additional keywords (and with terms associated with those keywords).
- *Evaluation*: Use earlier (document-level) TREC genomics test sets.
- *Dependencies*: Builds on WP1.
- *Uses*: Use thesaurus-biased language models as implemented (on top of Lemur) by [4]. Optimization methods described in [20, 39, 41, 60].
- *Risk*: Estimation of language models associated with keywords too expensive. *Remedy*: Use parsimonious models.

WP4 Modeling knowledge and structure. Combine ideas from WP2 and WP3 so that keyword-based evidence can be used (alongside content and structural information) at the element level. Pursue both a generative and discriminative approach to combining the evidence.

- *Output*: Models for combining heterogeneous element-level evidence.
- *Evaluation*: Test set for the Elsevier corpus from WP2.
- *Dependencies*: Builds on tools and models from WP2 and WP3. Depends on test set from WP2.
- *Uses*: Optimization methods described in [20, 39, 41, 60].
- *Risk*: Estimation too complex. *Remedy*: Reduce the number of tags in the Elsevier corpus by grouping them together. *Risk*: Resulting system too slow for interactive use. *Remedy*: Move computationally expensive steps off-line, and, if necessary, use approximations.

WP5 Evaluating content-based retrieval using knowledge and structure. Deploy models from WP4 to support simulated task-type interactive experiments and multiple locations. We choose for multiple locations both to be able to repeat the experiment with different subjects and to make sure that we cover a reasonable number of areas of the Elsevier corpus (computer science, biomedicine, food informatics).

- *Output*: Questionnaires plus analysis.
- *Evaluation*: Analysis of the interviews
- *Dependencies*: Depends on WP4.
- *Uses*: Interfaces from XMLFind/Wikiii [52]
- *Risk*: As for all complex software projects, underestimation of development time is a major risk. *Remedy*: Address by employing agile development practices with short iterations, systematic software design and rigorous testing.

WP6 Thesis writing and wrap-up.

- *Output*: Thesis manuscript. Consolidate code.
- *Dependencies*: Depends on all other WPs. WP2–WP5 are all expected to yield a workshop paper (or conference poster) and a conference or journal paper; these will form the backbone of the thesis, together with an extension of the survey written as part of WP1.
- *Risk*: Not enough material written or published during WP2–WP5. *Remedy*: There is a strict go/no-go decision towards the end of yr 1; at that time, the PhD student must have demonstrated sufficient writing skills and sufficient output.

c. Training and Education

Training and education are aimed at developing scientific expertise, and acquiring professional competencies. With respect to the scientific expertise, the Ph.D. student should become able to fully understand, critically analyze, and contribute to research at the frontiers of science. We address these issues both informally and formally. At an informal level, IvI's ILPS group provides a stimulating intellectual climate, with a range of world-class experts working on related projects (see the section on *embedding* for more details), with regular visitors (both academic and industrial), and with regular events (such as multiple seminars, small-scale workshops, national and international events).

More formally, supervision and training at the host institute are governed by a number of instruments. The Ph.D. student has two supervisors, and supervisors and student together draw up a training and education plan, which is revised (if necessary) annually. Our students follow graduate courses within the national research schools and at international summer schools. A working visit to a foreign university or research institute is a key ingredient of the plan. The training and education plan will also cover more general professional competencies (see below).

Towards the end of his first year, the Ph.D. student writes a detailed proposal for his thesis research and submits it to the supervisors and, after their approval, to an independent institute-wide doctoral committee that monitors Ph.D. students' progress on an annual basis. These formal progress evaluations are themselves subject to monitoring by external experts.

We actively work to equip our students to function well in the professional environment of a university or research institute. We encourage the development of academic leadership skills by involving Ph.D. students in various aspects of academic life, such as refereeing, project management, proposal writing, seminar/workshop organization, teaching, etc.

We encourage our Ph.D. students to complete an internship at relevant non-academic organization, so as to obtain valuable experience that bridges the gap between academic research and active applications of what has been learned. For the Ph.D. student on this project an internship at a major search engine (with scientific literature search facilities) or a major publisher (such as Elsevier) will be arranged.

The applicants are avid proponents and practitioners of an "open door" environment in which informal meetings occur naturally and frequently. In addition, the local research team members will have formal project meetings on a weekly basis, discussing both the research and training aspects, including mundane practical issues. Our attitude toward Ph.D. students is based on doing collaborative research where the initiative will gradually shift from the supervisors to the students, thus avoiding a strict (and, in our opinion counterproductive) dichotomy between supervisor and student.

8 Expected Use of Instrumentation

8.a. Resources

The project will build on UvA's Lemur-based language modeling tools.

For evaluation purposes we will use the following collections and setups:

- "Early" INEX collections (2002–2005); see [25]
- TREC genomics (2006); this edition provides a set of queries, a document collection of full-text biomedical articles, and relevance assessments. Relevance was measured at the document and aspect level. For our experiments, we use the judgments at the document level and those at the aspect level. See [59]
- Elsevier's corpus; 2.6GB; ∼50,000 full length articles covering computer science, biomedicine, and food informatics, with semantically informed document structure as well as keyword annotations. The corpus may be shared with other research groups and INEX for comparative research purposes.

8.b. Hardware

We do not request any additional storage or processing equipment.

9 Literature

Five most relevant publications by applicants

[1] C. Caracciolo and M. de Rijke. Generating and retrieving text segments for focused access to scientific documents. In M. Lalmas, A. MacFarlane, S. Rüger, A. Tombros, T. Tsikrika, and A. Yavlinsky, editors, *Advances in Information Retrieval: Proceedings 28th European Conference on IR Research (ECIR 2006)*, volume 3936 of *LNCS*, pages 350–361. Springer, April 2006.

[2] J. Kamps, M. Marx, M. de Rijke, and B. Sigurbjörnsson. Articulating information needs in xml query languages. *ACM Trans. Inf. Syst.*, 24(4):407–436, 2006. ISSN 1046-8188.

[3] J. Kircz. Creation driven marketing: Integrating metadata into the production process. *New Library World*, 108 (11/12):552–560, 2007.

[4] E. Meij and M. de Rijke. Thesaurus-based feedback to support mixed search and browsing environments. In *11th European Conference on Research and Advanced Technology for Digital Libraries*, pages 247–258, 2007.

[5] B. Sigurbjörnsson, J. Kamps, and M. de Rijke. Processing content-oriented XPath queries. In *Proceedings of the Thirteenth ACM Conference on Information and Knowledge Management (CIKM 2004)*, pages 371–380, 2004.

Other references

[6] ACM. Digital library, 2008. `http://www.acm.org/dl/`.

[7] K. Balog, L. Azzopardi, and M. de Rijke. Formal models for expert finding in enterprise corpora. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 43–50, New York, NY, USA, 2006. ACM Press.

[8] K. Balog, T. Bogers, L. Azzopardi, M. de Rijke, and A. van den Bosch. Broad expertise retrieval in sparse data environments. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 551–558, New York, NY, USA, 2007. ACM Press.

[9] M. J. Bates. What is browsingreally? a model drawing from behavioural science research. *Information Research*, 12(4), 2007. Paper 330. Available at `http://InformationR.net/ir/12-4/paper330.html`.

[10] P. Brusilovsky and M. T. Maybury. From adaptive hypermedia to the adaptive web. *Commun. ACM*, 45(5):30–33, May 2002.

[11] P. Brusilovsky, C. Karagiannidis, and D. Sampson. Layered evaluation of adaptive learning systems. *Int. J. Cont. Engineering Education and Lifelong Learning*, 14(4/5), November 2004.

[12] C. Caracciolo. *Topic Driven Access to Scientific Handbooks*. PhD thesis, Informatics Institute, University of Amsterdam, 2008. To appear.

[13] CiteSeer. CiteSeer Scientific Literature Digital Library, 2005. `http://citeseer.ist.psu.edu//`.

[14] M. Collins. Discriminative training methods for Hidden Markov Models: theory and experiments with the perceptron algorithm. In *Proceedings EMNLP 2002*, pages 1–8, 2002.

[15] P. De Bra, P. Brusilovsky, and G.-J. Houben. Adaptive hypermedia: from systems to framework. *ACM Computing Surveys*, 31(4es), 1999.

[16] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. John Wiley & Sons, Inc., 2001.

[17] S. Elbassuoni, J. Luxenburger, and G. Weikum. Adaptive personalization of web search. In K. Rodden, I. Ruthven, and R. W. White, editors, *SigIR 2007 workshop on Web Information Seeking and Interaction*, 2007.

[18] D. Ellis. Modelling the information seeking patterns of engineers and research scientists in an industrial environment. *Journal of Documentation*, 53(4):384–403, 1997.

[19] D. Ellis, D. Cox, and K. Hall. A comparison of the information seeking patterns of researchers in the physical and social sciences. *Journal of Documentation*, 49(4):356–369, 1993.

[20] J. Gao, H. Qi, X. Xia, and J.-Y. Nie. Linear discriminant model for information retrieval. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 290–297, New York, NY, USA, 2005. ACM.

[21] Google Scholar. Google Scholar, 2005. `http://scholar.google.com/`.

[22] W. D. Gravey. *Communication: The Essence of Science*. Pergamon Press, 1979.

[23] J. R. Herskovic, L. Y. Tanaka, W. Hersh, and E. V. Bernstam. A Day in the Life of PubMed: Analysis of a Typical Day's Query Log. *J Am Med Inform Assoc*, 14(2):212–220, 2007.

[24] H. Holz, K. Hofmann, and C. Reed. Unobtrusive user modeling for adaptive web-based systems. *International Journal of Pattern Recognition and Artificial Intelligence*, 21(2):225–244, 2007.

[25] INEX. Initiative for the Evaluation of XML Retrieval, 2008. `http://inex.is.informatik.uni-duisburg.de`.

[26] P. Ingwersen and K. Jarvelin. *The Turn: Integration of Information Seeking and Retrieval in Context (The Information Retrieval Series)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.

[27] J. Kamps, M. de Rijke, and B. Sigurbjörnsson. Topic field selection and smoothing for XML retrieval. In A. de Vries, editor, *Proceedings DIR 2003*, 2003.

[28] J. Kamps, M. Marx, M. de Rijke, and B. Sigurbjörnsson. XML retrieval: what to retrieve? In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 409–410, New York, NY, USA, 2003. ACM.

[29] C. Karagiannidis and D. G. Sampson. Layered evaluation of adaptive applications and services. In *Adaptive Hypermedia and Adaptive Web-Based Systems: International Conference, AH 2000, Trento, Italy, August 2000. Proceedings*, pages 343+. Springer, 2000.

[30] J. G. Kircz. New practices for electronic publishing 2: New forms of the scientific paper. *Learned Publishing*, 15(1):27–32, 2002. URL: `http://www.learned-publishing.org`.

[31] A. Kobsa. Generic user modeling systems. *User Modeling and User-Adapted Interaction*, 11(1):49–63, March 2001.

[32] T. Koch, A. Ardö, and K. Golub. Browsing and searching behavior in the renardus web service a study based on log analysis. In *JCDL '04*, pages 378–378, 2004.

[33] B. Larsen, S. Malik, and T. Tombros. The interaactive track at INEX 2005. In *Advances in XML Information Retrieval Evaluation: Fourth Workshop of the INitative for the Evaluation of XML Retrieval (INEX 2005)*, pages 398–410, 2006.

[34] M. Lehtonen, N. Pharo, and A. Trotman. A taxonomy for xml retrieval use cases. In *Comparative Evaluation of XML Information Retrieval Systems*, pages 413–422, August 2007.

[35] K. Matia, L. N. Amaral, M. Luwel, H. Moed, and H. Stanley. Scaling phenomena in the growth dynamics of scientific output. *Journal of the American Society for Information Science and Technology*, 56(9):893–902, 2005.

[36] E. Meij and M. de Rijke. Integrating conceptual knowledge into relevance models: A model and estimation method. In *International Conference on the Theory of Information Retrieval (ICTIR 2007)*. Alma Mater Series, October 2007.

[37] MeSH. Medical subject headings, 2008. `http://www.nlm.nih.gov/mesh/`.

[38] D. Metzler and W. B. Croft. A markov random field model for term dependencies. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 472–479, New York, NY, USA, 2005. ACM.

[39] D. A. Metzler. Automatic feature selection in the markov random field model for information retrieval. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 253–262, New York, NY, USA, 2007. ACM.

[40] M. J. Muller and S. Kuhn. Participatory design. *Commun. ACM*, 36(6):24–28, June 1993.

[41] R. Nallapati. Discriminative models for information retrieval. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 64–71, New York, NY, USA, 2004. Acm.

[42] Z. Nie, Y. Zhang, J.-R. Wen, and W.-Y. Ma. Object-level ranking: bringing order to web objects. In *WWW '05: Proceedings of the 14th international conference on World Wide Web*, pages 567–574, New York, NY, USA, 2005. ACM Press.

[43] P. Ogilvie and J. Callan. Hierarchical language models for xml component retrieval. In *Advances in XML Information Retrieval, Third International Workshop of the Initiative for the Evaluation of XML Retrieval*, pages 224–237, 2005.

[44] A. Paramythis and S. Weibelzahl. A decomposition model for the layered evaluation of interactive adaptive systems. In *User Modeling 2005*, pages 438–442. Springer, 2005.

[45] B. Piwowarski and P. Gallinari. A Bayesian network for XML information retrieval: Searching and learning with the INEX collection. *Information Retrieval*, 8(4):655–681, December 2005.

[46] *Report on the Royal Society Scientific Information Conference*, 1948. Royal Society.

[47] T. Saracevic. The stratified model of information retrieval interaction: Extension and applications. *Proceedings of the American Society for Information Science*, 34(2):313–27, 1997.

[48] ScienceDirect. Digital library, 2008. `http://www.sciencedirect.com/`.

[49] Scopus. Digital library, 2005. `http://www.scopus.com/`.

[50] H. Sharp, Y. Rogers, and J. Preece. *Interaction Design: Beyond Human Computer Interaction*. Wiley, March 2007.

[51] X. Shen, B. Tan, and C. Zhai. Implicit user modeling for personalized search. In *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 824–831, New York, NY, USA, 2005. ACM Press.

[52] B. Sigurbjörnsson. *Focused Information Access Using XML Element Retrieval*. PhD thesis, Informatics Institute, University of Amsterdam, 2006.

[53] B. Sigurbjörnsson, J. Kamps, and M. de Rijke. An element-based approch to XML retrieval. In N. Fuhr and S. Malik, editors, *Proceedings INEX 2003*, pages 19–26, 2004.

[54] B. Sigurbjörnsson, J. Kamps, and M. de Rijke. Focused access to Wikipedia. In *Proceedings DIR-2006*, 2006.

[55] SpringerLink. Digital library, 2008. `http://www.springerlink.com/`.

[56] D. E. Stokes. *Pasteur's Quadrant: Basic Science and Technological Innovation*. Brookings Institution, 1996.

[57] K. F. Tan, M. Wing, N. Revell, G. Marsden, C. Baldwin, R. MacIntyre, A. Apps, K. D. Eason, and S. Promfett. Facts and myths of browsing and searching in a digital library. In *ECDL '98*, pages 669–670, 1998.

[58] Thomson. The Web of Science, 2005. `http://www.isinet.com/products/citation/wos/`.

[59] TREC Genomics. TREC Genomics Track, 2008. `http://ir.ohsu.edu/genomics/`.

[60] R. Yan and A. Hauptmann. Query expansion using probabilistic local feedback with application to multimedia retrieval. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 361–370, New York, NY, USA, 2007. ACM.

[61] Y. Zhang and J. Koren. Efficient bayesian hierarchical user modeling for recommendation system. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 47–54, New York, NY, USA, 2007. ACM Press.

[62] W. Zhou, C. Yu, N. Smalheiser, V. Torvik, and J. Hong. Knowledge-intensive conceptual retrieval and passage extraction of biomedical literature. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 655–662, New York, NY, USA, 2007. ACM.

10 Requested Budget

`<Omitted>`