promise

PROMISE

Participative Research labOratory for Multimedia
and Multilingual Information Systems Evaluation

FP7 ICT 2009.4.3, Intelligent Information Management

Deliverable 4.1
First Report on Alternative Evaluation Methodology

Version 1.1, 31st August 2011

SEVENTH FRAMEWORK
PROGRAMME

Network of Excellence co-funded by the 7th Framework Program of the European Commission, grant agreement no. 258191

Document Information

Deliverable number:	4.1
Deliverable title:	First Report on Alternative Evaluation Methodology
Delivery date:	31/08/2011
Lead contractor for this deliverable	UBER
Author(s):	Richard Berendsen (UvA), Giorgio Maria Di Nunzio (UNIPD), Maria Gäde (UBER), Jussi Karlgren (SICS), Mihai Lupu (IRF), Stefan Rietberger (ZHAW), Juliane Stiller (UBER)
Participant(s):	UBER, UvA, IRF, ZHAW
Workpackage:	WP4
Workpackage title:	Evaluation Metrics and Methodologies
Workpackage leader:	UvA
Dissemination Level:	PU – Public
Version:	1.0
Keywords:	Evaluation, PatOlympics, LogCLEF, CHiC2011

History of Versions

Version	Date	Status	Author (Partner)	Description/Approval Level
1.0	24/08/2011	Draft	UBER	Circulated to reviewers
1.1	31/08/2011	Final	UBER	Revised after partners' comments

Abstract

The first report on alternative evaluation methodology summarizes work done within the PROMISE environment and especially within Work package 4 - Evaluation Metrics and Methodologies.
The report outlines efforts to develop and support alternative, automated evaluation methodologies, with a special focus on generating ground truth from existing data sources like Log files or annotations.

Events like LogCLEF 2011, PatOlympics 2011 or the CHiC2011 workshop are presented and reviewed on their impact on the three main uses case domains.

Table of Contents

Executive Summary

Complex multimedia and multilingual information systems require alternative and realistic evaluation methodologies according to predetermined use cases. PROMISE wants to improve current evaluation processes addressing the heterogeneity of users and diversity of information access systems. Several research projects are ongoing or have been completed focusing on the design of appropriate use cases and the corresponding evaluation of system performance and effectiveness.

This first report on alternative evaluation methodology summarizes work done within PROMISE and especially within Work package 4 - Evaluation Metrics and Methodologies.

Pointing out the need for alternative evaluation methodologies, different approaches for the development of concrete evaluation tasks and procedures are introduced.

Several research projects analyzing log files information to generate ground truth are explained and discussed in this report. Initial work dealing with the generation of relevance assessments derived from annotations and collections as well as plans for future work in this research area are presented.

Events like LogCLEF 2011, PatOlympics 2011 or the CHiC2011 workshop are reviewed on their impact on the three main uses case domains.

1 Introduction

One of the goals defined for PROMISE is the development and support of new methods and advanced metrics for more realistic evaluation procedures. Overcoming current limitations in IR evaluation is at least a 2-step process. First of all it is necessary to realize that user needs and system design are not independent factors dealing with complex multilingual and multimedia information systems. Therefore it is important to understand who the users accessing our systems are, their needs and goals, how they are searching or interacting and of course what they expect to find. As a following step it needs to be investigated how traditional measures and metrics can be enriched for effective IR evaluation according to the use case domains.

The main objective of work package 4 is to develop, analyze and ground novel methods for the evaluation of multimedia and multilingual information systems. This happens through the establishment of an evaluation infrastructure increasing efficiency and automation. For this deliverable two tasks are relevant within WP4:

- Task 4.1 – Generating Ground Truth from Log Files
- Task 4.2 – Generating Ground Truth from Collections and Annotations

For PROMISE, three main use case domains have been identified and serve as framework for the projects described in this report:

- **Unlocking culture**: deals with information access to cultural heritage material held in large-scale digital libraries comprising libraries, archives, museum, and audio-visual archives.
- **Search for innovation**: deals with patent search and its peculiar requirements to seek out standardized method and framework for evaluating different tools for the IP.
- **Visual clinical decision support**: deals with visual information connected with text in the radiology domain in order to provide retrieval and access mechanisms able to jointly exploit textual and visual features.

Other use case domains or use cases are also discussed if they work on alternative evaluation approaches.

Much work has already been conducted within the context of The Cross-Language Evaluation Forum (CLEF)[1]. In the past The Conference on Multilingual and Multimodal Information Access Evaluation promoted information retrieval system evaluation in monolingual and cross-language contexts. PROMISE will continue and improve the achievements of previous evaluation campaigns, providing an evaluation infrastructure which takes academic and industrial factors into account.

The report is organized as follows: Chapter 2 introduces the need for alternative evaluation methodologies as well as first approaches developed within the PROMISE environment. In

[1] http://www.clef-campaign.org/

Chapters 3 and 4 various research projects dealing with the analysis of log files information as well as collections and annotations are discussed. Especially the benefit of generating ground truth from these data sources is highlighted. General evaluation approaches such as PatOlympics2011 or CHiC2011 that help shaping new evaluation pathways are described in Chapter 5. We conclude with an outlook on future work and possible directions for further research projects dealing with the implementation of alternative evaluation methodologies.

2 Evaluation on the Move – Alternatives for the Cranfield Paradigm

The predominant evaluation paradigm in information retrieval to date is the so-called Cranfield paradigm as defined in [Voorhees 2002]. It aims to evaluate retrieval system performance by abstracting the problem of effective retrieval from operational variables. In the context of an information retrieval application, these variables are basically the environment of the system; namely data, configuration and user interaction. While it has been shown that upholding the paradigm in evaluations provides comparable and useful results, it should also be noted that industry relevant evaluation considers IR system performance as only one of many important factors. Voorhees deemed user-based evaluation "extremely expensive and difficult to do correctly". When taking a different focus, recent work [Braschler et al. 2006] has shown that approaches addressing user perception more directly are feasible in terms of effort and expressiveness.

Other approaches or methods for generating ground truth need to be considered and improved. From the industrial perspective, not only system performance but also user acceptance, usability criteria such as personalization and internationalization are becoming more and more important.

Future research and resulting evaluation methodologies and metrics cannot only focus on one aspect but need to take several aspects into account. In Section 2.1 and 2.2 two alternative evaluation proposals are presented: a use case-based and an application level evaluation. The chapter concludes with a description and discussion of the tasks in WP4 that involve alternative methodologies especially the generation of ground truth.

2.1 Goal-oriented Evaluation / Use Cases and Scenarios

One of the success criteria for a successful evaluation of an information access solution is the ability to predict subsequent take-up of the solution in practice. The connection between benchmarking and take-up is confounded by a large number of variables, which may be difficult to model, and the final quality of the complete system may hinge crucially on something completely different than the variables measured by benchmarking of its components. The informed choice of computational tools must be made on basis of their effect on the end usefulness of the system. If the effect of selecting a component which performs better in benchmarks cannot be measured in practice, it will be difficult to convince a commercial system designer to invest any effort in the improvements or in making the effort to find a solution other than the one developed in-house or included in the software framework otherwise in use.

Many approaches have been proposed to increase the realism of laboratory-based studies, not least within the interactive track of CLEF [Oard et al 2004, Clough et al 2008, Karlgren & Gonzalo 2010]. However, to translate the results from a lab study to the realities of fielded use is a non-trivial challenge, and frequently the hypotheses and the model which underly the experimental setting have not been made explicit. Recent approaches have addressed this through formulating an explicit task target context for user studies [e.g. Borlund 2000, Hansen 2005, Byström and Hansen 2005].

One methodological solution recently proposed by the CHORUS coordination action in deliverable 3.4 [Karlgren et al. 2009] and by this present project in deliverable 2.1 [Karlgren et al. 2011] is to develop *use cases*, which enable academic research efforts to vector their innovation and evaluation towards realistic usage scenarios. The point of a use case methodology in this case is to provide a bridge between on the one hand requirements analyses which describe known or hypothesized user needs and usage scenarios and on the other system components which can be benchmarked using traditional laboratory-based information access methodologies. The validation of the requirements analyses and the usage scenarios can be done by specialists in human-computer interaction or by user studies specialists, using any methodology which holds water for description of human behavior or commercial impact of systems -- the benchmarking can then be tailored to those requirements.

Here, a validated use case with clear and explicit hypotheses of usage goals and linked to evaluation benchmarks will be a much more convincing argument than a benchmark alone. Current and coming work in the PROMISE network of excellence will be geared towards validation and extension of the project-internal use cases to fit industrial and commercial stakeholders. The CHiC2011 workshop will investigate existing use cases within the cultural heritage domain with the aim to formulate future evaluation tasks (see Section 5.3).

2.2 From System to Application Evaluation

If our aim is to improve *industry relevance* of our work, we need to provide evaluation methods, which measure effects on the most important industry target: the users' (= clients`) perception. The system-based view by itself has its use for IR system implementers who directly profit from traditional IR evaluation research. Current market research of enterprise search [Andrews 2010] indicates few key players in the implementers' market with the majority of sales (and thus applications) being covered by affordable solutions by Google and Microsoft. However, the proportion of system implementers (building IR systems) to application providers (using IR systems in applications) is strongly skewed towards the latter. Corporations are interested in evaluating their enterprise search application's effectiveness and currently lack the means to correctly assess it.

The aforementioned observations have led us to believe that *evaluation on the application level is required* in addition to Cranfield-style system evaluation to provide a qualified assessment of retrieval performance for actual users. In Section 5.2 we describe two such studies that have been done. While they have been done before the PROMISE project started, we describe the main outcomes in this deliverable. They are large-scale studies of operational search engines. They evaluate systems from a user's perspective. And they combine many aspects of the systems, condensing them in four evaluation measures.

2.3 Generating Ground Truth from Log Files and Annotations

Traditional evaluation tasks often use human relevance assessments which are obviously expensive, time consuming and often biased. Making use of implicit or explicit signals from log file information or annotations enables more automated evaluation of information systems. Within WP4, two tasks are concerned with generating ground truth from rich data sources:

Task 4.1 – Generating Ground Truth from Log Files covers a wide range of log file analysis approaches: "The task will investigate the use of transaction log files for inferring relevance assessments. It will focus on rich logs that have more than just clicks (e.g, purchases) and on automatically inferring test-sessions (as opposed to individual queries) that can be used for (repeatable) evaluation experiments."

Task 4.2 - Generating Ground Truth from Collections and Annotations, which has just started, will deal with implicit and explicit annotations as relevance assessments: "This task will use (implicit or explicit) document annotations as relevance assessments. Explicit examples include keywords or categories assigned to documents in domain-specific collections. Implicit examples include labels assigned to linked-to documents, labels assigned to documents that a given document is grouped together with in resource sharing venues such as citeulike.org; another example is provided by patent search where ground truth can be generated based on different prior searches performed by human experts in the process of applying for, granting or opposing a patent."

The first information retrieval task that comes to mind when we talk about relevance assessments is ad hoc search. Ad hoc search refers to keyword search, perhaps the most common type of search nowadays. Search engines are typically evaluated by counting how many relevant documents they return for a set of queries, and how high they rank these documents. A relevance assessment states for a document and a query if the document is relevant to the query. Producing relevance assessments for ad hoc search tasks is labour-intensive and time consuming. Generating relevance assessments from log files has two main benefits: (i) it brings automation in the process of evaluating IR systems, and (ii) it involves real end users in the process. Work done in this field will be discussed in Section 3.

Two important research fields have evolved in the last years, query log analysis on the one hand and transaction or click stream log analysis on the other hand. Query log analysis is not necessarily only concerned with evaluating information access systems, another import aspect is the investigation of user behaviour by analyzing trends. Understanding what types of queries there are is an important part of query log analysis, and some of our efforts are directed to it. If automatic classification can be done with enough reliability to draw interesting and valid conclusions we believe this is a way of generating ground truth from log files.

An example of an automatic evaluation method that uses annotations is discussed in Section 4. Here patent retrieval is the task and patent citations are used as relevance assessments. These citations are withheld from the evaluated search engines.

If ground truth can be generated from log files or annotations to evaluate search engines, it should come as no surprise that such ground truth could be used to improve search engines. Research conducted to evaluate or optimize search engines cannot be regarded separately but is directly connected to each other.

The following two chapters provide an overview of relevant work done within PROMISE as well as research projects that address either alternative log file analysis approaches or use explicit and implicit signals from annotations or collections.

3 Ground Truth and Evaluation Approaches involving Log File Information

Various research projects deal with the analysis of log file data for system evaluation. Initiatives like LogCLEF provide log files for shared and comparative analysis and also address inherent limitations and challenges. For LogCLEF 2011, two research teams of Humboldt Universität zu Berlin and CELI s.r.l. provided ground truth in the form of annotations for a small subset of the TEL query logs (see Section 3.1). Other studies focus on the enrichment of logs (section 3.2) or domain specific query analysis (Section 3.3).

3.1 LogCLEF2011

The interactions between the user and an information access system can be analyzed and studied to gather user preferences and to learn what the user likes the most, and to use this information to personalize the presentation of results. Search logs are a means to study user information needs and preferences. The literature of log analysis of information systems shows a wide variety of approaches to learn user preferences by looking at implicit or explicit interactions [Agosti et al, 2011]. However, there has always been a lack of availability and use of log data for research experiments, which makes the verifiability and repeatability of experiments very limited. It is very difficult to find two research works on the same dataset unless by the same author, or where at least one of the authors worked for a commercial search engine company. This is not only a question of the same data source, but also a problem of using the same period of time for the analysis if the analysis has to be comparable with other works.

A first attempt to release a collection of log data with the aim of verifiability and repeatability was done within the Cross-Language Evaluation Forum (CLEF) [Agosti et al. 2010] in 2009 in the LogCLEF[2]track. Since 2000, CLEF promotes research in multilingual information access by supporting the development of tools for testing, tuning and evaluation of information retrieval systems operating on European languages in both monolingual and cross-language contexts, and creating test-suites of reusable data which can be employed by system developers for benchmarking purposes. LogCLEF is an evaluation initiative for the analysis of queries and other logged activities used as an expression of user behaviour [Di Nunzio et al., 2011, Mandl et al. 2011]. An important long-term aim of the LogCLEF initiative is to

[2] http://ims.dei.unipd.it/websites/LogCLEF/Overview.html

stimulate research on user behaviour in multilingual environments and promote standard evaluation collections of log data. In the three years of LogCLEF editions, different data sets have been distributed to the participants: search engine query and server logs from the Portuguese search engine Tumba! and from the German EduServer (Deutscher Bildungsserver: DBS); digital library systems query and server logs from The European Library (TEL); and Web search engine query logs of the Sogou Chinese search engine. Table 1 summarizes the log resources and the relative sizes.

Table 1. Log file resources at LogCLEF

Year	Origin	Size	Type
2009	Tumba!	350,000 queries	Query log
2009	TEL	1,900,000 records	Query and activity log
2010	TEL	760,000 records	Query and activity log
2010	TEL	1.5 GB (zipped)	Web server log
2010	DBS	5 GB	Web server log
2011	TEL	950,000 records	Query and activity log
2011	Sogou	730 MB (zipped)	Query log

In each edition of LogCLEF, participants are required to:

- Process the complete logs;
- Make publicly available any resources created based on these logs;
- Find out interesting issues about the user behaviour as exhibited in the logs; and
- Submit results in a structured file.

The public distribution of the datasets as well as the results and the exchange of system components aim at creating a community in order to advance the state of the art in this research area. The LogCLEF 2011 Lab presented four different tasks which tackle some of the issues presented in this work:

- Language identification task: participants are required to recognize the actual language of the query submitted.
- Query classification: participants are required to annotate each query with a label which represents a category of interest.
- Success of a query: participants are required to study the trend of the success of a search. The success can be defined in terms of time spent on a page, number of clicked items, actions performed during the browsing of the result list.
- Query re-finding, when a user clicks an item following a search, and then later clicks on the same item via another search; Query refinement, when a user starts with a query and then the following queries in the same session are a generalization, specification, or shift of the original one.

Another important aim of LogCLEF is to distribute ground truth generated manually or automatically by participants themselves. In CLEF 2010 the research teams of Humboldt Universität zu Berlin and CELI s.r.l. prepared annotations for a small subset of the TEL query logs. The annotated data contain:

- manual annotations for 510 query records about query language and category of the query;
- automatic annotations for 100 query records about query language.

The LogCLEF organizers also provided an interface for query language and classification.

3.2 Europeana Click Stream Logger

Depending on the information one needs or wants to extract, log data can be enriched or customized. Purposeful enriched log data serves as basic input for alternative or application based evaluation methodologies.

For the Europeana[3] Portal, a customized log format was developed with a special focus on language sensitive features such as interface language change or the use of language facets. Clickstream logging is a logging approach, which enables to mine complex data in order to analyze user behavior. A "clickstream" is a series of actions or requests on the web site accompanied by information on the activity being performed. It allows the tracking of application state changes and therefore traces user behavior in a way that a traditional http transaction log is unable to.

For the Europeana clickstream logs (CSL), different activity types or states with a particular focus on multilingual access aspects are logged [Gäde et al. 2010]. These actions indicate a stream of user activities which can be categorized as follows:

Interface language-specific actions: The interface language or the change of the interface language is logged for each transaction. This would otherwise not appear in the http access log.

Search-related actions: All search-related activities including information about query terms, result numbers and distribution of results by language and country are logged. Filtering (e.g restricting by language, provider, date) and related searches (from an initial result list) are also logged.

Browse-related actions: For the http transaction log, a browsing activity (e.g. clicking on one of the images cycling across the Europeana homepage that are suggested as search entries) is the same as a search via term entry: in both cases requests are sent to the search engine. However, from a user interaction perspective, browsing and searching might point to different user intentions. The clickstream logger logs all browsing activities and their initial starting points (e.g. did the search originate from the cycling images, the suggested searches, the time line, the saved searches links or saved user tags).

Navigation-related actions: User paths through search results are logged here, for example, when a user moves away from Europeana by following a link from a detailed results page to the original object or when the user returns to the results list.

User management-related actions: This involves actions connected to user account creation, logging in and out and changing passwords as well as interactive features of Europeana, are logged here. This includes saving or removing social tags, searches or objects. Additionally, errors and requests on static pages are logged.

[3] http://www.europeana.eu/portal/

3.3 People Search

The study of domain-specific search engines and especially in-depth query log analysis can provide a different view on search strategies. The more we know about user behaviour and expectation, the better systems can be designed and evaluated with respect to user needs.

More and more research shows increased attention for vertical search engines. These domain specific engines may benefit from domain knowledge. People search has been the subject of many studies. About four percent of queries in general web search were estimated to contain person names [Spink et al, 2004]. Instead of issuing person name queries in a general web search engine, many people use so called people search engines like 123people.com, Spock.com or kgbpeople.com. Domain knowledge may concern the types of queries commonly issued, the type of objects that are being searched for, relations between objects in the collection expressed in meta data, and so on. Within the PROMISE project, we are studying how domain knowledge may be used to improve the search process in the context of people search. As a case study, we study a Dutch people search engine, its query logs, typical queries, its interface, and characteristics of search results it returns.

People search engines queries consist of person names**.** Weerkamp et al. [2011] study the query logs of a Dutch people search engine. Queries issued to it consist of a first name, a last name and an optional keyword. Contrasting people search with general web search; they find a much higher percentage of one query sessions and a low click through ratio. The latter finding may be explained by the fact that in the user interface of the people search engines it takes three clicks to actually leave the search result page and follow the out link that is registered in the logs. Social media profiles are the most clicked result type.

The types of queries issued are of interest because an engine may choose different retrieval strategies, or ways of presenting the results based on the type of query. Weerkamp et al. [2011] propose a taxonomy of people search queries that was inspired by typical patterns in the trend of search volume of queries. High profile queries are queries for people who are receiving much attention. They can be subdivided into event based queries and regular high-profile queries. When interest in a person spikes because of a recent event, it is typically an event based query. When interest is high all the time because of many events, or because somebody is an established celebrity or public person, we talk about typically regular high profile queries. An experiment was performed to classify queries automatically in the proposed taxonomy. Berendsen et al. [2011] analyze a classification experiment of people search queries in more detail. The main findings of both experiments are that it is easy to separate high profile queries from low profile queries, but that distinguishing event-based queries from regular high profile queries is much harder.

3.4 Relevance Information from Transaction Logs

As stated before log file information can serve as an alternative to human relevance assessments. Following we discuss two research lines that mine search logs for inferring relevance. The first approach, simulating queries, turns the search process upside down. Instead of starting with an information need, then formulating a query, matching documents to this query and ranking the documents, it starts with the question: Which documents are likely to be relevant to any query? Then from such a document a query is generated by sampling terms from the document. This is where log files come in. Log files are used to inform the sampling distribution. When the query is generated, this query-document pair is used as a relevance assessment, fully automating the evaluation of search engines. Another interesting aspect of this work is the ground truth relevance assessments used to validate the proposed evaluation methodology. Rather than using manually crafted relevance assessments purchase decisions obtained from transaction logs are used. This work has been presented at CLEF 2010.

The second approach, interleaving ranked lists, has a tradition that goes back to 2003, when Thorsten Joachims coined the idea of merging the ranked lists produced by two different search engines, recording the clicks, and counting which search engine contributed most clicked search results. Exactly how to merge the search result lists to make the evaluation fair has been a subject of study in recent years, and several improvements over the original merging algorithm have been found. We discuss a recent approach that shows very strong results.

Notable in the work we discuss is how the different merging strategies are evaluated. A learning to rank dataset of queries and feature vectors for returned documents for these queries released by Microsoft is used. Relevance assessments were available, but no user clicks. In recent years, several user models that predict clicking behaviour have been developed. In this process, transaction logs have been used extensively to validate the models. To obtain clicks for the learning to rank dataset, a state of the art user model is implemented. It is adapted to use the graded relevance assessments available.

Interleaving ranked lists to evaluate search engines is an idea that has found uptake in industry already, with several papers co-authored by search engine operators. Another aspect of the approach is that it makes use of end user feedback – however noisy and implicit — to inform evaluation rather than expert judgements. A limitation is that it only serves to judge the quality of a ranked list of results. For example, user behaviour in image search, where images are lined up side to side and top to bottom is less well understood. Other aspects of evaluation, such as performance speed, are not considered.

3.4.1 Simulating Queries

Simulated queries have been compared to manually created queries for information retrieval [Azzopardi et al. 2007, Tague et al. 1981]. Reproducing absolute evaluation scores through simulation has been found to be challenging, as absolute scores will change with, e.g., recall base. However, reproducing exact retrieval scores is not essential to developing a useful simulator when we wish to rank retrieval systems by their performance. Following this argument, the aim is to make a simulator that allows us to identify the *best performing retrieval* system.

Our work is similar to [Azzopardi et al. 2007] as our goal is to assess simulators for retrieval evaluation. However, we focus on relative performance instead of absolute scores as we argue that this is a more feasible and useful goal. Instead of comparing simulators to explicit judgments for known-item queries, we compare our approaches to a large number of purchase-query pairs that are derived from implicit judgments obtained from a transaction log. We apply and extend simulation strategies developed in [Azzopardi et al, 2007], and compare these to strategies that take characteristics of logged queries into account.

3.4.2 Implicit Feedback from Click Data – Interleaving Ranked Lists

Hofmann et al. [2011] reviewed different approaches for evaluating ranks using click data as implicit feedback. Since none of these methods provided satisfying results, they propose a new probabilistic method, which allows considering different rankers that are difficult to compare for other methods use before.

Radlinski et al. [2008] investigate evaluation methods based on usage metrics such as document clicks, query reformulation or time spend on a page derived from click through data. The study compares two methodologies; one assuming that retrieval quality affects directly user click behavior ("absolute metrics") and another one based on balanced interleaving ranks ("Paired Comparison Tests"). They could not find a significant relationship between retrieval quality and implicit feedback whereas the interleaving tests showed good results concerning the quality judgment.

Both studies highlighted the advantages of implicit feedback but also remarked that the accurate interpretation is crucial for the quality, generalization and reproducibility of results. Future work needs to focus on overcoming and minimizing inherent limitations and challenges.

3.5 Improving Algorithms using Transaction Logs

In this Section we discuss two research activities with the aim to generate ground truth from log files not so much for evaluation purposes as for improving retrieval performance. These two aims can be flip sides of the same coin: if it is possible to generate ground truth in the form of good result rankings in an automatic way, the same algorithms should be useful for optimizing a search engine. Research into any of these two fields is therefore bound to be useful for the other field.

3.5.1 TREC Sessions

The University of Amsterdam (UvA) participated in the TREC Session Track 2011. The objective of the track was to investigate if previous queries in a session can be used to improve retrieval for the current query. Participants were invited to submit runs in four conditions:

1. Not using information from previous queries,
2. Using only the previous queries,
3. Using also results seen for previous queries,
4. Using also clicks on previous search result pages.

For the track the organizers had created test topics, most of which were explorative in nature. Test subjects (researchers, mostly) were invited to choose topic on subjects they knew well or found interesting, and search for them on a search engine from which subsequently click data was recorded.

Team UvA chose to derive relative relevance assessments from the clicks in condition 4. This was done using ao. the CLICK > SKIP ABOVE strategy developed by Joachims et al [2005]. The idea behind this strategy is that people typically scan a search result list from top to bottom. They click when they think the result is relevant. For each recorded click the strategy then assumes that skipped (not clicked) results above clicked results are less relevant than the clicked document. These relative relevance assessments were then used to optimize a linear combination of several features of documents.

3.5.2 People Search

A well recognized problem for handling people search queries is ambiguity of person names. According to US Census data, in a sample of about seven million US inhabitants, ninety percent of the people shared only about ninety thousand unique names. In the past, several tasks have been organized around the problem of named entity disambiguation in search results for person name queries; the WePS (Web People Search) tasks [Artiles et al. 2007, 2009, 2010]. If this problem could be solved, a search engine could greatly improve its search result presentation: documents could be grouped around the individuals they refer to, simplifying the task for an end user who wants to find all information about a particular person.

The search results in the WePS setting were obtained using a general purpose web search engine. In the PROMISE project we revisit this problem in a different setting: queries issued to a specialized people search engine. Its result presentation is different: results are grouped according to their type: Facebook profiles, LinkedIn profiles, Google search results, Yahoo! search results, and so on. Because of this, social media profile profiles are much more common than in the search results of a general purpose search engine.

We are interested in extracting ground truth from log files of the people search engine. Ground truth in this case is a mined signal from the log files that helps in clustering the search results. An example idea is that when a query has a peak in its search volume history, queries at this time are more likely to be related to the same person than queries outside of the peak. Therefore, if two documents are clicked on for this query during such a peak it increases the probability that they refer to the same person. Our work on this is ongoing.

4 Ground Truth from Collections and Annotations

Work within task 4.2 - Generating Ground Truth from Collections and Annotations started in month 9 and builds on the experiences and results from task 4.1. Goal of this task is to find ways of using annotations generated by a critical mass of users to assess relevance. These assessments can be derived from implicit or explicit annotations.

Structured data on the Internet is available in larger and larger quantities. The Linked Open Data cloud, Wikipedia and the Open Directory (http://www.dmoz.org) are examples of huge annotated corpora created in large part by volunteers. On many social media platforms, people routinely tag or rate web pages, video fragments, images and so on. There has been research into using this user-generated content for evaluation of search engines.

Beitzel et al. [2003a, 2003b] sample queries from a log file that exactly match entries in the Open Directory 11, a huge directory of web pages in a large taxonomy, maintained by people. Then they assume that these Open Directory entries are relevant for the queries and evaluate IR systems with these relevance assessments. This technique is promising for evaluating known item searches.

4.1 Work Plan for Task 4.2

Task 4.2 involves the identification of explicit and implicit annotations and their use for meaningful relevance assessment. Following are some examples of annotation variants in collections, which will be completed according to our findings:

Explicit annotations:

- keywords or categories assigned to documents in domain-specific collections
  - tags, controlled vocabularies
- votes / groupings for patents

Implicit annotations:

- labels assigned to linked-to documents
- grouping of documents in a resource-sharing venue such as cite-u-like.org
- ground truth can be generated based on different prior searches performed by human experts in the process of applying for, granting or opposing a paten

Several collections and associated annotations such as The European Library (TEL), Bibsonomy or Flickr have already been studied and could serve as input or test corpus for future research. Especially within the patent search domain, available and extended high quality metadata is a very useful source for automated relevance assessments. IRF provides a test corpus for research called MAREC[4].

For PROMISE task 4.2, in-depth studies of identified and selected collections will be conducted in order to evaluate the impact on new evaluation methods using automated relevance assessments. In the following Sections, we shortly discuss two examples of evaluation using annotations, one in the domain of patent retrieval and another in a different kind of information retrieval task that has come to be known as ‘wikifying’. The two examples share one principle: information present in the data is withheld from the algorithms, which have to reproduce it.

4.2 Leveraging Patent Citations

In the patent retrieval domain, relevance assessments can be derived from patent citations listed on patent search reports. The methodology of extracting these ground truths is described in detail in [Graf et al. 2008]. Every patent search reports lists the most relevant prior patents which are supposed to be closely related to each other, the present patent accruing from the prior one. The automatically extracted citations can serve as relevance assessments for prior art searches.

[4] http://www.ir-facility.org/prototypes/marec/statistics

For the CLEF-IP campaigns[5] in 2009 and 2010, this methodology was developed further by extending the list of references extracted from the prior search reports [Piroi et al. 2011]. In addition to extracting the patents referenced in the search report, citations which are listed in the family members to the topic patents were added to the list. This leads to an increase of culminating citations by a factor of 7. In figure 4.1, the methodology of extracting direct and related citations is visualized.

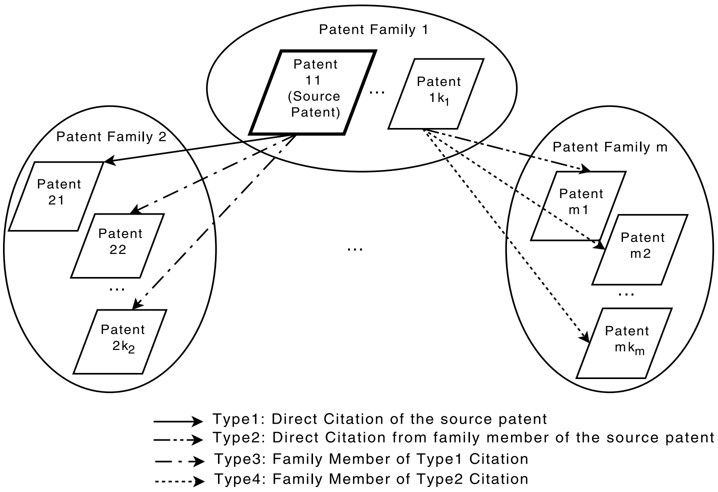
Patent Family 1
Patent 11 (Source Patent)
...
Patent 1k$_1$
Patent Family 2
Patent 21
Patent 22
...
Patent 2k$_2$
...
Patent Family m
Patent m1
Patent m2
...
Patent mk$_m$
Type1: Direct Citation of the source patent
Type2: Direct Citation from family member of the source patent
Type3: Family Member of Type1 Citation
Type4: Family Member of Type2 Citation

Fig. 4.1 Patent citation extension used in CLEF–IP

4.3 Leveraging Links to Wikipedia

Wikifying is the task of linking words or phrases that need more explanation to Wikipedia pages. A common way of evaluating wikifying algorithms is to apply them to random Wikipedia pages that have their links removed and counting how many links are found back.

However, there may be problems with this approach. If a different set of random pages is used, results may vary. More importantly, people may disagree about which phrases are candidates for linking. There exist many trivial links in Wikipedia, such as year, country, etc. which are actively rejected by human assessors [Huang et al. 2009]. Will algorithms that have been developed while being evaluated on random general domain Wikipedia pages generalize well to domain specific test collections? We now discuss some research that shows the answer is at least sometimes ‘no’.

[He et al. 2011] studied the problem of generating links to Wikipedia articles in radiology reports: data from a medical domain. They found that two state of the art methods, Wikify! and Wikipedia Miner, performed much worse than previously reported on data from the general domain. The main bottleneck is anchor detection: deciding which phrases need to be linked. The major cause identified for the change in performance is that medical phrases typically have a more complex semantic structure than Wikipedia concepts. An algorithm was developed that makes use of regularities in phrase structure and it achieved substantial

[5] http://www.ir-facility.org/clef-ip

improvements over the two state of the art methods. This is a good example of how properties of a specific domain may be exploited to improve retrieval performance.

This Section gave a first insight in possibilities to extract ground truth from implicit or explicit annotations. The main challenge is to understand the conditions of the use case domain the ground truth is generated for. The two examples above showed that domain expertise is needed to interpret the annotations correctly. It is possible to use approaches from the general domain but expertise is needed to translate this in to the specialized fields. For future work, this approach will be developed further.

5 Other Alternative Evaluation Approaches

Additionally to the research projects mentioned above, several initiatives and industrial activities aim at moving towards alternative or domain specific evaluation approaches. Focusing on the main uses cases defined for PROMISE several projects are planned to advance the evaluation of complex information systems. Following we present a selection of three different ways of dealing with specific challenges and requirements facing IR evaluation tasks that aim at moving forward an advanced evaluation infrastructure.

5.1 PatOlympics 2011

For the patent search, it is especially important to include as much context information as possible, finding all relevant publications to a special topic. Evaluation efforts for this use case domain are addressed by the CLEF-IP[6] track. Another initiative dealing with evaluation of patent retrieval systems is the PatOlympics[7]. The 2011 edition of the PatOlympics took place on the morning of the IRF Symposium, on June 7th, 2011. As plenary session every registered member (this year's IRFS had around 60 registrants) could attend and see how IP professionals used the systems provided by the IR teams in the two sports: ChemAthlon and CrossLingual Retrieving.

The PatOlympics evaluation metrics and methodology are subjected to a series of opposing forces because the audience is not necessarily IR researchers familiar with such acronyms as nDCG, MAP, PRES, etc. and at the same time, we need to maintain at least a minimum level of reliability of the results and fairness in the competition. We detail the causes and the consequences of these opposing forces in tables 2 and 3, along with the actions taken to meet them.

[6] http://www.ir-facility.org/clef-ip

[7] http://www.ir-facility.org/events/irf-symposium/irf-symposium-2011/patolympics

Table 2: Force IP: Requirements and Actions

Requirement	Action
Users not familiar with **metrics popular in the IR community and without the time to learn a new field**	Used a very simple metric: sum of Recall@200 over all topics, expressed simply as "Number of relevant documents found"
Evaluators (i.e. *Referees***) do not have the time to create full relevance judgements for their topics prior to the event**	The system allows a referee to add relevance judgements for their topic at any time: before, during, or after the event. Scores are automatically recalculated when this happens.
The evaluation is live, therefore the process is intellectually complex and it cannot last very long	The event is executed in rounds, where each round is about 25 minutes and in total the event does not last for more than 3 hours
Force IP: Attendees and target audience from the IP professional domain	

Table 3: Force IR: Requirements and Actions

Requirement	Action
The scores must be comparable	The score is always computed based on the latest relevance judgements of the referee, even if a participant is no longer in the game due to the scheduling of the rounds
The time allocated to each topic for each team must be the same	Teams can only send in candidate relevant documents during the round when this topic is allocated to them. They cannot see the topic before and cannot send in results afterwards.
The relevance judgements must be consistent throughout the event	As we involve human evaluators, they may indeed become more aware of their own topic as they see results. There is nothing that can be done against this. However, the issue is mitigated by the fact that we have more than 2 topics per game, and therefore each team will be equally [dis]advantaged by being the first, or last team a referee works with
Force IR: The evaluation must follow best practices in the IR Evaluation field	

The event is scheduled to run in rounds. At each round, a referee sits down at a team's table, presents the request for information, and works with the team to answer this request for information, while at the same time telling the system, independently of the team, which documents he found relevant in the current round, if any.

The PatOlympics were extremely well received by all participants, be they referees or teams. It was explicitly requested to the organizers to try and organize it again. Despite the success, the event, and in particular the evaluation framework behind it needs further work. The PatOlympics system is currently performing the very basic operations (receiving input, calculating scores, displaying results). It requires a better design and needs to include management tools in order to deal with any unforeseen situations. Currently, such tasks are done mostly manually and fed into the back-end database directly.

Furthermore, it might be interesting to create a totally remote version of the event, where all participants would take part remotely. We had some experience this year, when one of the participants, CMU – which turned out to be a winner in ChemAthlon – participated remotely and interacted with the referees via Skype. This is a considerable improvement over the 2010 campaign, where we also had a remote participant, but in which case the connection did not actually work and therefore that team could not actually participate, to the disappointment of the team but of the referees and IRFS attendees as well. Relying on the internet connection of the event location is always a point of failure and a fully distributed event may mitigate that.

5.2 Black Box Evaluation of Enterprise Search Applications

As mentioned before, PROMISE also wants to examine the use of operational systems.Two studies [Braschler et al. 2006, Braschler et al. 2009] were conducted on Swiss and German enterprise search portals by a set of corporate and academic partners including ZHAW. The goal was to determine the current state of site search implementations in Switzerland and Germany for mid-sized to large enterprises and public institutions. Furthermore, an evaluation grid was to be created, which could be used for future evaluations. Site search functionality is treated as an *application* consisting of an IR system, its configuration and the information made accessible. This application is evaluated as a whole.

The main evaluation tool in these studies is an *evaluation grid*. It contains a large number of mostly independent weighted criteria. Interdependent criteria are summed up into one overall criterion and given fractional weights (i.e. 3 sub-criteria being weighted at 1/3 and summed up for a total weight of 1). Usually, a single independent criterion is weighted as 1. Criteria belong to tests, which in turn belong to one of four main categories in the studies: *search index, query/document matching, user interaction, search result.*

The evaluation was carried out using as many different systems and criteria as possible. Measurements were taken in a simple manner, e.g. counting hits or "true/false" The results for each criterion per system were entered into the aforementioned evaluation grid. This allowed consistent ranking of systems and measurement of average, typical performance. Tests were mainly done manually, with a few tests having been carried out automatically by the corporate partner Eurospider. Testers were students and employees of the various project partners. About 4-6 hours of effort was required per site. All tests were run externally; *treating the application as a black box*, and no modifications of the web sites in response to testing was possible.

The second study introduces a *Google baseline test* to be able to compare site search testing results to a well known reference. It was carried out fully for three sites, using Google web search with a site restriction akin to “site:company.de”. The Google baseline scored slightly ahead of the German sites’ median scores. It did, however, not deliver the top result for any main category. As had been expected because of user experience with the application, Google excelled in the user interaction category. Google web search is not optimized for the specific enterprise’s data or its users and updates can not be controlled. In some cases, updates may even be intentionally suppressed by robot exclusions.

Testers had also been tasked to rate 2 separate “soft criteria”: give an *overall impression* of the tested site with a score ranging from 1 (inacceptable) to 10 (excellent) and a *fun factor* of 0 (no fun), 1 (okay) or 2 (fun), including other unstructured comments. The general impression score had an average of 4.84 and a median of 5. This is remarkably close to the absolute median scores of three out of four main categories.

Some basic recommendations for enterprise search implementers conclude the study reports:

- Keep index clean, complete and current;
- Maintain metadata (correct time stamps, titles, etc.);
- Follow accepted standards for user interaction (avoid unexpected behaviour);
- Control search results (result list from user’s perspective!); and
- Serve common information needs well.

It has been shown that a comprehensive relative evaluation of search applications with only external access to the application is feasible. Additionally, it is evident that user satisfaction and actual search performance are affected by many more factors than only a suitably well implemented IR system.

The studies drew a lot of interest from corporations whose web sites had been tested. The corporations’ decision makers (e.g. CTOs, IT managers, etc.) were all invited to presentations of the work and many of them did show up. This stands in strong contrast to regular events of this sort, where only academic peers and few familiar corporate partners of the presenting institutions are present.

5.3 CHiC 2011 – Cultural Heritage in CLEF

The “Unlocking Culture” domain deals with information access to cultural heritage material held in large-scale digital libraries comprising libraries, archives, museum, and audio-visual archives. Different to the other use case domains, no standard evaluation procedure for Cultural Heritage use cases exists therefore no standard requirements are defined yet.

The CHiC 2011[8] workshop to be held during CLEF 2011 aims at moving towards a systematic and large-scale evaluation of cultural heritage digital libraries and information access systems and helping to shape a possible roadmap for it.

Digital libraries and other information systems that access *Cultural Heritage* (CH) materials are becoming increasingly complex. They must often manage a diverse range of content from different CH institutions – such as libraries, museums, written and audiovisual archives – and have to provide access to them in a unified and coherent way. The content from CH institutions is often multilingual and multimedia (e.g. text, photographs, images, audio recordings, and videos), usually described with metadata in multiple formats and of different levels of complexity. CH institutions have different approaches to managing information and serve diverse user communities, often with specialized needs. This makes the meaning of "search and browse" quite different for users of a library or archive and non-specialist users may not be able to successfully retrieve relevant results or may be disoriented by the kind of results they obtain. Much effort is being placed on designing and developing effective search systems and tackling issues such as user interfaces, interoperability and metadata enrichment.

Interaction patterns of users with CH information systems do not represent clear separated and isolated use cases but should be understood as components which complement and alternate with each other thus representing possible sequences of user interactions with a CH information system.

Invited talks by Jaap Kamps, Johan Oomen and Christos Papatheodorou will address further challenges and possibilities of alternative evaluation activities. Complementary participant are asked to bring in statements dealing with the following topics:

- Use cases, evaluation needs, and best practices coming from field experience in the cultural heritage institutions;
- Evaluation perspectives, frameworks, and approaches in the digital library and digital curation fields;
- Synergies and relationships between large-scale evaluation campaigns and CH evaluation.

The objective of this workshop is to review existing use cases in the CH domain and their translation into potential retrieval and evaluation scenarios that can be used as benchmarks for evaluating CH information access systems. The overall goals are:

- To establish what makes searching in the cultural heritage domain distinct from other domains
- To gather existing use cases for multilingual information access in the CH domain.
- To review existing evaluation resources studies within the CH domain.
- To propose appropriate methodologies for evaluating multilingual information access to CH resources.

[8] http://www.promise-noe.eu/chic-2011/home

- To define multiple concrete evaluation tasks modeled on IR evaluation initiatives such as CLEF, TREC or INEX

Based on the input and outcomes of this workshop it is planned to organize an appropriate evaluation task for next year.

6 Conclusions

The research introduced in this first report on alternative evaluation methodology can be seen as a starting point for continuous improvement of evaluation methodologies, methods and metrics. Especially in the field of log file analysis, a variety of studies have been conducted with relevant results for the PROMISE project. The evaluation approaches discussed here depend and benefit partially on work conducted in WP6 – Evaluation Activities. Especially activities related to the use case domains can be reused for the improvement of information retrieval evaluation of domain specific systems.

In general, a clear movement from traditional, system-oriented to a more user-oriented and use case-based evaluation set-up, which considers industrial needs can be observed.

Further work is planned within Task 4.2 - Generating Ground Truth from Collections and Annotations. As stated above, this task has just started leveraging alternative resources for evaluation purposes. Future work will concentrate on the identification of suitable test collections and associated annotations.

The outcomes of LogCLEF2011 and CHiC2011 will serve as input for future work as well as for the final report on alternative evaluation methodology due in August 2012.

References

[Agosti et al. 2011] M. Agosti, N. Ferro, C. Peters, M. de Rijke, and A. Smeaton. Multilingual and Multimodal Information Access Evaluation: International Conference of the Cross-Language Evaluation Forum, CLEF 2010, Padua, Italy, September 20-23, 2010, Proceedings, volume 6360. Springer-Verlag New York Inc, 2010.

[Agosti et al. 2010] M. Agosti, F. Crivellari, and G. Di Nunzio. Web log analysis: a review of a decade of studies about information acquisition, inspection and interpretation of user interaction. Data Mining and Knowledge Discovery, pages 1–34, 2011.

[Andrews 2010] W. Andrews. Market Scope for Enterprise Search. http://www.gartner.com/, Document ID: G00206087. 22nd November 2010.

[Artiles et al. 2010] J. Artiles, A. Borthwick, J. Gonzalo, S. Sekine, and E. Amig´o. WePS-3 Evaluation Campaign: Overview of the Web People Search Clustering and Attribute Extraction Tasks.

[Artiles et al. 2009] J. Artiles, J. Gonzalo, and S. Sekine. Weps 2 evaluation campaign: overview of the web people search clustering task. In 2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference, 2009.

[Artiles et al. 2007] J. Artiles, J. Gonzalo, and S. Sekine. The semeval-2007 weps evaluation: Establishing a benchmark for the web people search task. In Proceedings of the 4th International Workshop on Semantic Evaluations, pages 64–69. Association for Computational Linguistics, 2007.

[Azzopardi et al. 2007] L.Azzopardi, M. de Rijke, and K. Balog. Building simulated queries for known-item topics: an analysis using six European languages. In SIGIR ’07, pages 455–462, New York, NY, USA, 2007. ACM.

[Bailey et al. 2008] P. Bailey, N. Craswell, I. Soboroff, P. Thomas, A. P. de Vries, and E. Yilmaz. Relevance assessment: are judges exchange- able and does it matter. In SIGIR ’08, pages 667–674, 2008.

[Beitzel et al. 2003a] S. Beitzel, E. Jensen, A. Chowdhury, D. Grossman, and O. Frieder. Using manually-built web directories for automatic evaluation of known-item retrieval. In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, pages 373–374. ACM, 2003.

[Beitzel et al. 2003b] S. Beitzel, E. Jensen, A. Chowdhury, and D. Grossman. Using titles and category names from editor-driven taxonomies for automatic evaluation. In Proceedings of the twelfth international conference on Information and knowledge management, pages 17–23. ACM, 2003.

[Berendsen et al. 2011] R. Berendsen, B. Kovachev, E. Meij, M. de Rijke, and W. Weerkamp. Classifying queries submitted to a vertical search engine. Koblenz, 2011. ACM.

[Borlund 2009] Borlund, P. Interactive Information Retrieval in Digital Environments. Journal of the American Society for Information Science and Technology, 60: 1944–1945, 2009.

[Brascher et al. 2009] M. Braschler, B. Heuwing, T. Mandl, C. Womser-Hacker, J. Herget, P. Schäuble, J. Stuker. Evaluation der Suchfunktion deutscher Unternehmenswebsites. In: Wissensorganisation 09: “Wissen – Wissenschaft – Organisation” 12. Tagung der Deutschen ISKO (International Society for Knowledge Organization), 2009.

[Braschler et al. 2006] M. Braschler, J. Herget, J. Pfister, P. Schäuble, M. Steinbach, J. Stuker. Evaluation der Suchfunktion von Schweizer Unternehmens-Websites. In: Churer Schriften zur Informationswissenschaft, Schrift 12, 2006.

[Byström and Hansen 2005] Byström, K., Hansen,P. Conceptual Framework for Task in Information Studies. JASIST - Journal of the American Society for Information Science and Technology, 56 (10). pp. 1050-1061, 2005.

[Clough et al. 2008] Clough, P., Gonzalo, J., Karlgren, J., Barker, E., Artiles, J., Peinado, V. Large-scale interactive evaluation of multilingual information access systems: the iCLEF Flickr challenge. In: Workshop on Novel Methodologies for Evaluation in Information Retrieval, 30 March 2008, Glasgow, Scotland, 2008.

[Di Nunzio et al. 2011] G. Di Nunzio, J. Leveling and T. Mandl. Multilingual log analysis: Logclef.
Advances in Information Retrieval, pages 675–678, 2011.

[Gäde et al. 2010] M. Gäde, V. Petras and J. Stiller. Which Log for Which Information? Gathering Multilingual Data from Different Log File Types. In Proceedings of CLEF. 2010, 70-81.

[Graf et al. 2008] E. Graf, L. Azzopardi. A methodology for building a patent test collection for prior art search. In: Proceedings of the second international workshop on evaluating information access (EVIA), 2008.

[Hansen 2005] Hansen, P. Work Task-Oriented Studies of IS&R Processes. Developing Theoretical and Conceptual Frameworks to be applied for evaluation and design of tools and systems. In: Theories of Information Behaviour. ASIST Monograph Series . ASIST, Medford, NJ, USA, pp. 392-396, 2005.

[He et al. 2011] J. He, M. de Rijke, M. Sevenster, R. van Ommering, Y. Qian. Automatic Link Generation with Wikipedia: A Case Study in Annotating Radiology Reports, 20th ACM Conference on Information and Knowledge Management (CIKM 2011), Glasgow, ACM, October, 2011.

[Hofmann et al. 2011] K. Hofmann, S. Whiteson, M. de Rijke. A Probabilistic Method for Inferring Preferences from Clicks, 20th ACM Conference on Information and Knowledge Management (CIKM 2011), Glasgow, ACM, October, 2011

[Huang et al. 2009] W. C. Huang, A. Trotman, and S. Geva. The importance of manual assessment in link discovery. In SIGIR ’09, pages 698–699, New York, NY, USA, 2009.

[Joachims et al. 2005] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR ’05). ACM, 2005.

[Joachims et al. 2003] T. Joachims. Evaluating retrieval performance using click- through data. In J. Franke, G. Nakhaeizadeh, and I. Renz, editors, Text Mining, pages 79–96. Springer Verlag, 2003

[Karlgen et al. 2011] J. Karlgren, G. Eriksson, M. Frieseke, M. Gäde, P. Hansen, A. Järvelin, M. Lupu, H. Müller, V. Petras and J. Stiller. PROMISE Deliverable 2.1: Initial specification of the evaluation tasks. PROMISE Project Consortium, 2011.

[Karlgren and Gonzalo 2010] Karlgren, J. and Gonzalo, J. (2010). Interactive Image Retrieval. In: ImageCLEF - Experimental Evaluation in Visual Information Retrieval. The Information Retrieval Series, (32). Springer, pp. 117-138, 2010.

[Karlgren et al. 2009] J. Karlgren, M. Kauber, N. Boujemaa, R. Compañó, C. Dosch, J. Geurts, H. Gouraud, P. King, J. Köhler, P. van der Linden, R. Ortgies, Å. Rudström, N. Sebe. CHORUS Deliverable 3.4: Vision Document. Chorus Project Consortium, 2009.

[Mandl et al. 2011] T. Mandl, M. Agosti, G. Di Nunzio, A. Yeh, I. Mani, C. Doran, and J. Schulz. Logclef 2009: the clef 2009 multilingual logfile analysis track overview. Multilingual Information Access Evaluation I. Text Retrieval Experiments, pages 508–517, 2011.

[Oard 2004] Oard, D., Gonzalo, J., Sanderson, M., Ostenero, L., Wang, J. Interactive Cross-Language Document Selection. Information Retrieval 7, 1-2, 205-228, 2004.

[Piroi et al. 2011] F. Piroi and V. Zenz, “Evaluation Information Retrieval in the Intellectual Propery Domain: The CLEF-IP Campaign” in Current Challenges in Patent Information Retrieval, Lupu et al (Eds), Springer Verlag, 2011

[Radlinski et al. 2008] F. Radlinski, M. Kurup, and T. Joachims. How does click-through data reflect retrieval quality? In CIKM ’08, pages43–52, 2008.

[Spink et al. 2004] A. Spink, B. Jansen, and J. Pedersen. Searching for people on web search engines. Journal of Documentation, 60(3):266–278, 2004.

[Tague et al. 1981] J.M. Tague and M.J. Nelson. Simulation of user judgments in bibliographic retrieval systems. SIGIR Forum, 16(1):66–71, 1981.

[Voorhees 2002] E. M. Voorhees. The philosophy of information retrieval evaluation. In: CLEF '01: Revised Papers from the Second Workshop of the Cross-Language Evaluation Forum on Evaluation of Cross-Language Information Retrieval Systems. Springer-Verlag, London, UK, pp. 355–370. 2002.

[Weerkamp et al. 2011] W. Weerkamp, B. Kovachev, R. Berendsen, E. Meij, K. Balog, and M. de Rijke. People searching for people: Analysis of a people search engine log. In 34th Annual International ACM SIGIR Conference (SIGIR 2011), Beijing, 2011.