PROMISE

Participative Research labOratory for Multimedia and Multilingual Information Systems

FP7 ICT 2009.4.3, Intelligent Information Management
www.promise-noe.eu

Deliverable 4.3.
Final Report on Alternative Evaluation Methodology

Version 1.1, September 2012

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

PROMISE
Participative Research labOratory for Multimedia and Multilingual Information Systems

SEVENTH FRAMEWORK PROGRAMME

Document Information

Deliverable number:	4.3
Deliverable title:	Final Report on Alternative Evaluation Methodology
Delivery date:	31/08/2012
Lead contractor for this deliverable	UBER
Author(s):	Richard Berendsen, Martin Braschler, Maria Gäde, Michael Kleineberg, Mihai Lupu, Vivien Petras, Stefan Rietberger
Participant(s):	UBER, ZHAW, UvA, TUW
Workpackage:	WP4
Workpackage title:	Evaluation Metrics and Methodologies
Workpackage leader:	UvA
Dissemination Level:	PU – Public
Version:	1.1
Keywords:	Evaluation Methodologies, Use Case, CHIC2012, Evaluation, Use Cases, CLEF2012

History of Versions

Version	Date	Status	Author (Partner)	Description/Approval Level
1.0	27/08/2012	Draft	UBER	Circulated to all partners
1.1	04/09/2012	Final	UBER	Revised after partners' comments

Abstract

The final report on alternative evaluation methodology collects work done within the PROMISE project, especially within Work package 4 – Evaluation Metrics and Methodologies, as well as related research projects and initiatives.

The report summarizes efforts and achievements focussing on alternative, automated or improved evaluation methodologies.

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

Related events or initiatives like PatOlympics 2012 or the CHIC2012 lab to be held within CLEF2012 are discussed and reviewed on their impact on the use case domains.

Table of Contents

Document Information 3
Abstract 3
Table of Contents 5
Executive Summary 6
1 Introduction 7
2 Ground Truth from Annotations and Collections 8
2.1 Explicit Annotations 9
2.1.1 Using Controlled Vocabularies to estimate Relevance 10
2.1.2 Non-Controlled vocabularies 15
2.2 Creating Simulated Queries from Keywords 17
2.3 Pseudo test collections for learning to rank scientific articles 19
2.4 Simulated Relevance Assessments from Annotations 20
2.5 Explicit annotations in Search for Innovation 21
3 Ground Truth from Log files 23
3.1 Result Disambiguation in Web People Search 24
3.2 Interpreting clicks as relevance feedback 24
4 Alternative Retrieval Scenarios and Evaluation Metrics 25
4.1 CHIC2012 – Cultural Heritage in CLEF 25
4.1.1 CHIC: Diversity Task 25
4.1.2 CHIC: Semantic Enrichment Task 26
4.2 Application-Centric Black Box Evaluation / User Perception 28
4.3 Expert finding and expert profiling 31
4.4 Reputation Management 31
4.5 PatOlympics 2012 32
5 Future Work and Conclusions 32
References 34

Executive Summary

Complex multimedia and multilingual information systems require alternative and realistic evaluation methodologies according to predetermined use cases. The PROMISE project 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.

The final report on alternative evaluation methodology takes up work presented in the first report on alternative evaluation methodology as well as novel evaluation approaches within PROMISE and especially efforts within Work Package 4 - Evaluation Metrics and Methodologies.

Research dealing with the generation of relevance assessments derived from annotations and collections is presented. The theoretical framework for two experiments developed within Task 4.2 - Generating Ground Truth from Collections and Annotations are described and discussed.

Furthermore events like PatOlympics 2012 or the CHiC2012 lab are reviewed on their impact on the uses case domains.

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

1 Introduction

Overcoming limitation in IR evaluation through the development and improvement of new methods and appropriate metrics for evaluation procedures is one of the goals defined for PROMISE. Traditional system-centred evaluation focused on quantitative measures for effectiveness of search interactions (information retrieval tests where success is measured as finding relevant documents). Although the Cranfield paradigm [Cleverdon 1997, Voorhees 2002] dominates the experimental set-ups and has proven its value in system comparisons, increasingly different evaluation methodologies are sought. While the laboratory setting ensures the control of the experimental design, it abstracts from real systems and posits fixed assumptions about use scenarios. The evaluation of operational systems and challenges with respect to the assumptions about user needs (stable), user-system interaction (single search input) and relevance (binary and objectively measurable) caused the development of new methods and new measures in recent years. Within the information retrieval community, a general call for more user engagement, more realistic search scenarios and the involvement of or combination with user-centric methods can be observed [Ingwersen & Järvelin 2005].

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 activities like the CLEF initiative are also discussed with regard to alternative evaluation approaches.

The deliverable includes efforts done on WP4 Evaluation Metrics and Methodologies, especially within the following tasks:

- Task 4.2 – Generating Ground Truth from Collections and Annotations
- Task 4.3 – Alternative Retrieval Scenarios and Evaluation Metrics

The deliverable mainly focuses on work done within Task 4.2 – Generating Ground Truth from Collections and Annotations while research done within Task 4.1 – Generating Ground Truth from Log Files has already been discussed in the first deliverable on alternative evaluation methodology will be only briefly touched in this

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

report. Work related to Task 4.3 – Alternative Retrieval Scenarios and Evaluation Metrics will be retreated in D4.4 – Report on Operational Systems as Experimental Platforms where the results for the Guerilla Evaluation Campaign will be discussed in detail.

The report is organized as follows: Chapter 2 presents a detailed research review on automatic estimations of relevance using annotations in document collections before four different approaches for generating ground truth from annotations and collections in the PROMISE project are presented. Chapter 3 looks back at task 4.1 (generating ground truth from log files) because more research was generated during the reporting period. Chapter 4 introduces alternative evaluation tasks and measures that are used in the PROMISE use case domains and other research, while chapter 5 provides a conclusion.

2 Ground Truth from Annotations and Collections

In information retrieval, the matching of information needs to relevant documents is a key function. Ultimately, end users of search engines judge the relevance of retrieved documents to their information needs. In this section we explore if annotations produced by end users, e.g. through tagging, can be used to approximate this notion of relevance. For example, documents annotated with a tag 'java' may be assumed to be relevant to a query 'Java'. We distinguish relevance assessments that are estimated automatically from different clues (generated ground truth) from relevance assessments based on intellectual judgements of retrieved documents (editorial ground truth). The underlying idea behind task 4.2 was to generate collections of documents that are estimated to be relevant (relevance assessments) based on automatically generated corpora using clues like annotations in document instead of using expensive intellectual assessments for judging a document's relevance based on an information need. Ground truth obtained in this way can be used for several purposes:

Benchmarking

Benchmarking measures the ranking of a set of retrieval algorithms by their performance on the generated ground truth. The quality of generated ground truth is typically estimated by comparing the retrieval scores using the ground truth to retrieval scores on an editorial test collection [Chowdhury & Soboroff 2002, Beitzel et al. 2003c, 2003b, 2003a, Hawking et al. 2004, Amitay et al. 2004].

Training

The generated ground truth is used as a tuning or training set. The quality of this training material can be evaluated using retrieval performance on editorial ground

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

truth. It may be compared to performance when editorial ground truth is used for training [Asadi et al. 2011, Berendsen et al. 2012a].

In section 2.1 we provide an overview of the literature on generating ground truth using annotated corpora. We make a distinction between annotations from controlled vocabularies, e.g. Medical Subject Headings (MeSH) [French et al. 2001] and non-controlled vocabularies, e.g. folksonomy data [Trant 2009]. Before we dive into approaches to generate ground truth from annotations, we treat approaches that use annotations to improve retrieval algorithms directly, e.g. through query expansion [French et al. 2001, Hersh et al. 2000, Taghva et al. 1999], or through re-ranking [Hotho et al. 2006, Kamps 2004]. There is an interesting interplay between these approaches and using annotations for ground truth. When we use generated ground truth for benchmarking, we must be careful that it is not biased towards any of the systems being benchmarked [Jensen et al. 2007]. This danger is exacerbated if some of the benchmarked systems use the same annotations that were used to generate ground truth to retrieve documents that are then assessed using the ground truth. When we use generated ground truth for training, we have to make sure that the difficulty of the problem posed by the generated ground truth is representative of the difficulty of solving real test problems: we usually model real test problems with queries from an editorial test collection. It is an interesting research question whether systems that use the same annotations from which ground truth is mined can be trained as successfully as systems that do not use these annotations. The following studies show that in many cases, ground truth can be mined from annotations such that both benchmarking and training are successfully facilitated.

2.1 Explicit Annotations

The evaluation of large-scale information retrieval systems needs to avoid the time-consuming and cost-intensive process of manual relevance judgments. Therefore, alternative approaches of automated or semi-automated evaluation methods are explored. Ali & Sufyan Beg [2011] presented an overview of different web search evaluation methods and emphasized the importance of automated evaluation, although they concluded that the significance of human evaluation cannot be neglected. A similar conclusion is drawn by Soboroff et al. [2001] in the attempt to evaluate information retrieval systems by replacing human relevance judgments with a randomly selected mapping of documents to topics considered as pseudo-relevance-judgments. Such a minimal-effort approach can only illustrate relative system performance and not replace a true test collection for measuring effectiveness. In another review Hersh & Kim [2006] suggested that some fundamental other approaches are needed if manual assessments were to be replaced.

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

A new field of interest consists of approaches concerned with the exploitation of pre-existing manual organizations of documents like explicit annotations as reviewed by Sanderson [2010]. Such human edited explicit annotations can be considered as forms of already given manual relevance assessments, meaning somebody already assigned keywords describing the content (therefore describing the relevance to an information need asking for this content). Annotations can be separated in controlled and non-controlled vocabularies. Examples of controlled vocabularies are thesauri, authority files and different kinds of classifications like traditional library classifications as well as ontologies, web site taxonomies, site maps or web directories. In this case, keywords or categories are chosen from a pre-defined and structured vocabulary to annotate documents. In contrast, terms of non-controlled vocabularies are pieced together in a flat structure from free tags by indexers or user groups. Relevance judgments using social tagging are exploiting folksonomy data from social bookmarking services.

2.1.1 Using Controlled Vocabularies to estimate Relevance

Traditional controlled vocabularies such as thesauri, library classifications or subject headings are manually built and, therefore, can be exploited to improve information retrieval performance or to simulate manual relevance judgments. Results have been mixed, however.

Using controlled vocabularies to improve retrieval performance

French et al. [2001] used the *Medical Subject Headings* (MeSH) as a controlled vocabulary for mapping user queries into an augmented query expansion. The results have shown that the expanded queries boost the collection selection performance by more than 25% and also outperform the original queries for document retrieval.

A similar approach is examined by Hersh et al. [21] using MeSH terms for thesaurus-based query expansion. In this study hierarchical query expansion showed up to 29.5% improvement. The conclusion is drawn that in general this thesaurus-based approach declines retrieval performance, at the same time the results improve in specific cases.

In a further study Hersh et al. [1996] compared two MEDLINE searching systems, the first one based on traditional Boolean searching on human-indexed thesaurus terms, and the second one based on natural language searching on phrases in abstract, title or indexing terms. The evaluation showed no significant differences between these two systems.

Taghva et al. [1999] described an automatic query expansion test using a collection-specific thesaurus in the Boolean-based environment BASISplus. The result for the

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

thesaurus relation type *preferred terms* did not show a significant improvement in recall and precision.

Harter & Cheng [1996] proposed the information retrieval technique *co-linked descriptors* to improve vocabulary selection for query terms using the thesaurus of ERIC descriptors. Co-linked descriptors are seen as an analogy to the bibliometric measure co-cited references. In general, the authors concluded that co-linked descriptors are an effective tool to generate useful search terms, but they have to be used in conjunction with an AND relation in order to produce reasonably high precision.

Iivonen & Sonnenwald [1998] proposed a model of search term selection to cover multiple discourses on the same topic. Six different discourses, which are referring to the ways of thinking and talking, are examined as a source of search terms such as controlled vocabularies, documents, domains, practices of indexing, search requests, databases, and search experience. The result has shown that users change discourses dynamically during the search term process and that controlled vocabularies are the preferred discourse.

Dolin et al. [1998] examined the use of classification schemes such as the *Library of Congress Classification* (LCC) to cluster heterogeneous information sources. Based on the information architecture *pharos* an amount of newsgroups, each considered as an individual collection, was automatically classified using the hierarchical classification schemes of LCC. The authors concluded that the proposed search technique is able to improve web search and is applicable to any digital text collection.

Humphrey [1999] proposed an automated document indexing approach using journal descriptors from databases such as MEDLINE. The advantages are seen in the fact that the document set does not depend on manual indexing. The conclusion was drawn that the most promising use of the rather general journal descriptors would be for refining and improving search results.

Shiri et al. [2002] compared thesaurus-based search interfaces of research-related and commercial web sites. Although the interface design features differ significantly the functionalities which most likely improve information retrieval are seen as an explicit thesauri search option, an understandable terminology and a hierarchical as well as an alphabetical term list.

In the context of human-computer interaction Beaulieu [1997] described interface functionalities for query expansion. The tension is emphasized between automatic vs. interactive query expansion, explicit vs. implicit use of thesauri, and document vs. query space. The latter is described as a shift from document-centered interface to one that takes the importance of query building into account.

Suomela & Kekäläinen [2005] examined query formulation with and without conceptual support such as the use of ontologies. The *Concept-based Information Retrieval Interface* (CIRI) produced a higher number of search terms, but did not

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

improve the three measures generalized precision, precision based on personal assessments, and generalized relative recall.

In the context of user interaction Joho et al. [2004] examined concept-based query expansion using the TREC test collection. It has been shown that the automatically generated concept hierarchies reduce the number of iterations and improve the finding of relevant items. This improvement of precision is considered most effective at the higher document levels.

Soergel [1997] explored the use of multilingual thesauri in text and speech retrieval. It is argued for the development of a common conceptual system as a reference point for all languages. Furthermore, it is emphasized that thesauri for a knowledge-based support of searching do not require that users are experts in thesauri and classification.

Kamps [2004] explored the use of controlled vocabularies from classifications for a new re-ranking strategy for initially retrieved documents. The tests with the *German Indexing and Retrieval Test Data Base* (GIRT) and the French *Amaryllis* collections demonstrated that the information retrieval effectiveness in domain-specific collections was significantly improved. Although, it is emphasized that re-ranking strategies cannot improve recall, but only precision.

The GIRT collection, which consists of reports and papers in the social sciences domain, was also used by Gey & Jiang [2000] to explore the exploitation of a multilingual thesaurus. While in general the result has shown that multilingual queries in English and German do not achieve the best performance, the conclusion was drawn that the exploitation of the GIRT thesaurus can more than double retrieval precision.

In related work, Petras [2005] and Petras et al. [2002] used the GIRT collection to test a thesaurus-based query expansion to disambiguate and translate search terms. Although the technique *Entry Vocabulary Modules* showed only minimal improvement over baseline retrieval, the combined machine translation with thesaurus matching achieved better results, in particular for individual queries.

Jin et al. [2002] explored the use of category labels from metadata such as topic information in XML format in a new language model to improve retrieval accuracy. For the comparison a *Text Retrieval Conference* test collection (TREC4) and automatically extracted labels via k-means clustering were tested. The results outperformed traditional language models and it is emphasized that the proposed approach can be applied for other types of metadata.

Using controlled vocabularies to generate ground truth

Since the creation of the hypertext structure within websites also involves manual organization, the taxonomies or topical classifications presented on many web sites can also be seen as a form of relevance assessment. The exploitation of the web

site taxonomies has been tested by Hamandas et al. [1997] to build ground truth for an image test collection, although this approach is seen as applicable for any kind of documents. The assessments were guided by the topical categories of the taxonomies to identify target areas of those pages that might contain sub-sets of images relevant to related queries. While this method is seen as restricted to the use for queries that reflect the topical classifications, the process for gathering relevance judgments was significantly sped up.

Another automated evaluation methodology are techniques using manually constructed web directories, such as the *Yahoo!Directory* or the *Open Directory Project* (ODP). In the context of interactive web search Bruza et al. [2000] compared the effectiveness of keyword search against browsing a human edited web directory under the assumption that the latter would lead to higher relevance of documents. The results did not reveal that directory-based search using *Yahoo!Directory* improved relevance over standard query-based web search using Google.

Chowdhury & Soboroff [2002] used the ODP for an automated construction of query-document pairs as a baseline for the evaluation of five web search engines. While the queries are mined from real query logs, the relevant documents are drawn from the human edited OPD entries. In this automatic evaluation approach it is assumed that web searches have a navigational intent.

Likewise, Beitzel et al. [2003c] considered web pages as relevant known-items for queries that match the editor-entered titles. In a large-scale system evaluation using the web directories ODP and *LookSmart* it has been shown that this automatic approach is not biased by the directory used, and has a moderately strong positive correlation to manual evaluations.

In a following examination Beitzel et al. [2003b] additionally assumed that all entries in the same leaf-level category of a web directory are also relevant. While the former result has been confirmed, it was found that the category-match has a weaker correlation with the manual evaluation than the title-match technique.

In a further publication Beitzel et al. [2003a] described the used measures like Pearson rank and Spearman rank in more detail and concluded that the stronger correlation in the title match evaluation is based on the used best-document style method with few retrieved documents in comparison with the weaker correlation in the category match evaluation which produced many pseudo-relevant documents. Nevertheless, all evaluation methods agreed which three of six search engines performed best and which three worst.

Haveliwala et al. [2002] examined the possibility of using the hierarchical order in the web directories *Yahoo!Directory* and ODP for relevance assessments related to web page similarity search. The assumption that a given document is on average more similar to other documents in the same class or in related classes is reflected in the definition of the *familial distance*, which operates with four possible values (0-3) from same class, through sibling classes (same superordinate class) and cousin classes

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

(superordinate classes with the same next superordinate class) to unrelated classes. While the fact is acknowledged that familial distance does not always hold, this evaluation strategy seems to sufficiently reflect the notion of similarity embodied in web directories for generating ground truth.

The distance in the ODP hierarchy is also used by Menczer [2003] to estimate precision and recall for web search engines in the context of semantic mapping. The similarity between two web pages is quantitatively measured in the relationships between content, link, and semantic topology. In contrast to Haveliwala et al. [2002], this approach can use any query, because the additional link similarity as well as the content similarity using term frequency enable bootstrapping a virtual relevant set of highly related pages to a few examples of pages with manual explicit annotations.

In the field of topical search evaluation Maguitman et al. [2010] explored the use of further properties of web directories for exploiting semantic similarity data. In addition to the hierarchical structure of ODP, many non-hierarchic components provide semantic relationships like “related” or “symbolic” edges to bridge the gap between different branches of classes, which is why a the web directory ODP as an ontology is considered not as a tree-based, but a graph-based similarity including cross-references. The results provided experimental evidence for support in effectiveness of automatic evaluation of topical retrieval systems.

Cecchini et al. [2011] examined a similar approach to the evaluation of topical search using topic semantic similarity data derived from ontologies like the ODP. In addition to traditional evaluation metrics the novel performance measures *semantic precision* and *semantic harmonic mean* are proposed. It has been shown that semantic measures provide the best techniques to find highly relevant documents via existing relationships between web pages. This framework is expected to evaluate also other information retrieval applications, like classification and clustering algorithms.

A semi-automatic evaluation framework is developed by Jensen et al. [2007], in which a small number of manual judgments are combined with pseudo-relevance judgments mined from web directories like ODP or *LookSmart*. In the dynamic environment of web search it is shown that the additional automatic judgments improved evaluation accuracy and reduced errors by half. It is suggested that the semi-automatic methods are applicable to different kinds of human-edited classifications such as a web directory, a corporate intranet directory or even a large collection of categorized bookmarks as long as such explicit annotations are not biased towards particular search services.

In the context of enterprise search Hawking et al. [2004] used site maps to assess relevance. Most organizational web sites include a site map as an overview of the information available and as an index for every single web page within the structure of the enterprise site. In general, such lists are maintained by the web site owner and developed by experts who are familiar with the enterprise, which is why site maps can be considered as reliable explicit annotations for documents. In this case

the documents are single located web pages and the authors emphasized that this evaluation strategy is only applicable to navigational search, in which the identified pages can be seen as known items.

Even traditional library classifications in electronic form in OPACs or digital libraries can be exploited for information retrieval tasks. Bosca & Dini [2009] proposed a novel approach for query enrichment and expansion using multilingual categories from library classification systems considered as natural document clustering. Although the examination is not primarily concerned with evaluation methods the developed *Word2Category* approach indicates further possibilities for exploiting explicit annotation for information retrieval systems.

Amitay et al. [2004] described an evaluation method using *Term Relevance Sets* (*Trels*) instead of relevant documents assessed by human judgments. Trels is based on a list of relevant onTopic and irrelevant offTopic terms to each query and examines the occurrences of these terms in the retrieved documents. It has been shown that the results are highly correlated with the well known TREC measures. Since the terms may be keywords, phrases or lexical affinities this approach can be considered as an example for using explicit annotations.

He et al. [2011a, 2011b] generate links to Wikipedia in narrative radiology reports. They leverage existing links within Wikipedia to train two state of the art algorithms but find that these algorithms do not generalize well to the highly domain specific medical radiology reports. They propose a new algorithm and outperform the two state of the art algorithms, both in identifying which terms in the reports should be linked as well as in identifying to which Wikipedia page they should be linked. We reported in more detail on this work in deliverable 4.1 [Berendsen et al. 2011b].

Building on Beitzel et al. [2003b] but extending that work with a query generation technique and pursuing the purpose of training a learning to rank system rather than the purpose of benchmarking retrieval algorithms, Berendsen et al. [2012a] generated a pseudo test collection in the digital libraries domain using curated annotations. This work is discussed in Section 2.3.

2.1.2 Non-Controlled vocabularies

In contrast to traditional web directories, explicit annotations or tags from user groups are not restricted to a pre-defined vocabulary, but reflect personal preferences and are indicators for the web user's interests. Trant [2009] reviewed the research related to social tagging data until 2007 and discussed the role of user tags in information retrieval.

In the context of personalized searching systems, Xu et al. [2008] proposed an automatic evaluation framework based on folksonomy data from social bookmarking services like *Del.icio.us* or *Dogear.* In comparison to categories of web directories like ODP the results demonstrated that social annotations are higher quality

descriptors of web pages. Under the condition that enough user-specific relevance judgment data are available, it has been shown that the exploitation of folksonomies significantly improve search quality.

In a related work, Zhou et al. [2008] combined the modeling of social annotations with a language modeling approach of information retrieval using a data sample collected from *Del.icio.us.* Therefore, the terms in annotation tags has been considered as additional content terms of documents to categorize their topics. The results demonstrated that the combined method outperformed traditional approaches, although the dominant web related topics in *Del.icio.us* are suggested to be biased.

For further exploitation of *Del.icio.us* dat,a Bao et al. [2007] proposed two novel algorithms, namely the *SocialSimRank* (SSR) to discover the latent semantic association between queries and the explicit annotations as well as *SocialPageRank* (SPR) to provide a static ranking from the perspective of the annotator.

In the context of web search Morrison [2008] compared the retrieval performance of tagging data from *Del.icio.us* against search engines and the web directory ODP. While the search engines generally showed higher precision, the *Del.icio.us* performed better than ODP. It is suggested that the folksonomy search results could be used to improve the retrieval performance of search engines.

Hotho et al. [2006] proposed the search algorithm *FolkRank* to exploit the structure of folksonomies for a personalized re-ranking. The conclusion is drawn that *FolkRank* provides best results for topical related elements in querying folksonomies.

Ramage et al. [2009] use tagging data from *Del.icio.us* for clustering web pages into semantic groups. K-means clustering and a novel clustering algorithm based on *latent Dirichlet allocation* are examined. The results demonstrated that the inclusion of tagging data improve cluster quality in comparison with page text alone.

Furthermore, Markines et al. [2009] evaluate semantic similarity measures for social tagging with a focus on the similarity among tags and resources using *WordNet* and ODP. The question of scalability was highlighted and the results have shown that measures based on collaborative aggregation of explicit annotations leads to the best performance.

In the context of web search personalization via *Del.icio.us* tagging data Vallet et al. [2010] proposed two novel personalization techniques to re-rank result lists. The first one is based on a vector space model using the concepts *tag inverse document frequency* and *tag inverse user frequency* and the second one is based on a probabilistic model using the Okapi BM25 ranking model. The result has been shown that these techniques outperform previous personalization approaches.

In a related work Noll & Meinel [2007] proposed a web search personalization approach exploiting tagging data to re-rank search results. While this approach

improved search results independent of the search engines, the personalization performed better for users which were not only broadly interested, but topical experts.

Another attempt to exploit user-generated data is examined by Liu et al. [2009] using *Wikipedia* disambiguation pages. In the context of diversity search a test collection is generated that includes documents with ambiguous topics, which are open to different interpretations and a broader range of relevance judgments. After sampling the queries with different levels of *Similarity of Intentions* (SI) it has been shown that this manual approach is feasible to construct a test collection for evaluating search result diversity, and it is considered that the human effort could be minimized by semi-automatic methods.

Asadi et al. [2011] use anchor texts in web pages as a source for query terms, and linked-to documents as potentially relevant documents. Their objective is not to generate a pseudo test collection for the purpose of benchmarking retrieval algorithms. Rather, they aim to train a learning to rank (LTR) algorithm on the generated ground truth. They compare LTR performance to a BM25 retrieval model. They also compare performance with the same LTR algorithm trained on the same editorial judgments that were used for testing.

Sidebar: Terminology in Describing Annotations

It should be noted that the authors make different use of terminology. For example, Harmandas et al. [1997] use “taxonomy” for a topical classification within a web site, in contrast, Jensen et al. [2007] mainly use “taxonomy” for a web directory of web sites, but also for corporate intranet directories or large collections of categorized bookmarks. On the other hand, there is a lack of distinction in the interchangeable use of the terms “annotation”, “keyword”, “category”, “tag” or “label”, even Xu et al. [2008] who drew a line between “category” (related to a structured classification) and “keyword” (related to a general topic) are confusing the terms. Likewise, the *Open Directory Project* is mostly considered as a classification but occasionally as an ontology. Additionally, the distinction of “implicit” and “explicit” annotations seems to be in question, when Haveliwala et al. [2002] use the phrase “implicit ordering information” referring to categories of a web directory.

2.2 Creating Simulated Queries from Keywords

Evaluation campaigns make use of a test collection including queries and a set of documents. The following experiment describes an alternative way of simulating queries, starting with a set of keyword combinations and relevant documents in order to create appropriate information needs and queries. It is the goal to investigate to what extent existing keywords can be leveraged to create simulated queries and collections.

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

The bilingual GIRT collection (GESIS) serves as a baseline for the experiment. GIRT contains more than 150,000 metadata records in German and English for publications in the social science domain. The GIRT collection is organized and manually indexed according to controlled vocabularies such as thesauri, subject headings and classification notations. Using a combination of keywords assigned to the documents, a sub-collection was created which contains all documents that are relevant for the Boolean expression (e.g. KW1 + KW2). The definition of concepts required keyword combinations that result in appropriate result sets (>12; <50 documents fulfilling the Boolean query). The query keywords were then removed from the sub-collection. Based on the document content, an appropriate information need was generated manually. A researcher read the documents and described what information need the documents would fulfil. In another step, the information need was then translated into a query. The identification of information needs and queries was done by independent users since the knowledge of keywords could bias the process. The assessors based their decision only on the titles and abstracts of all relevant documents for each keyword combination.

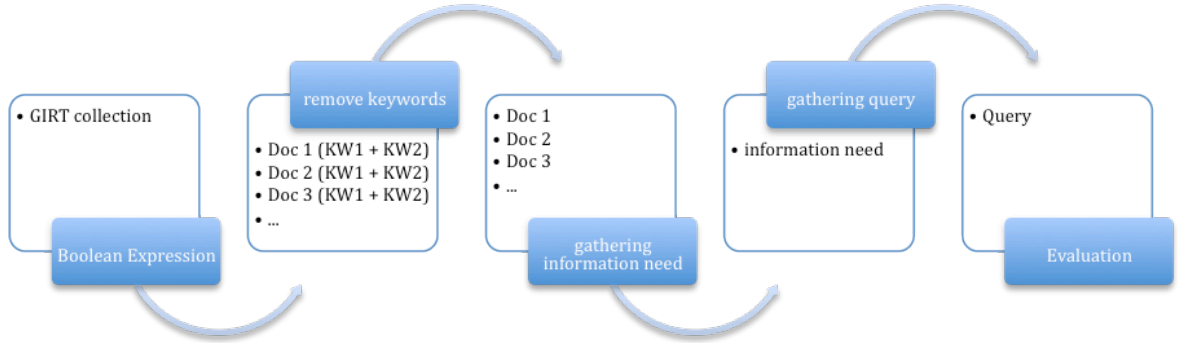

Figure 1: Overview Experiment I

The created ground truth contains 37 German matching information needs, queries and a collection of relevant documents without having to assess thousands of documents with respect to their relevance for the given information needs.

Comparing the keyword combinations with the created information needs showed an almost exact match with the keywords. After the transformation of information needs to queries a lower consistency was observed.

Keyword Combination	Information Need	Query
finanzielle Situation (KW1) geschlechtsspezifische Faktoren	Die Dokumente handeln vom Zusammenhang zwischen ökonomischer Situation und sozialen	Zusammenhang ökonomische Situation sozialer Hintergrund

(KW2)	Aspekten z.B.: Geschlecht, Familienstand, Alter, Herkunft etc.	
Computer (seit 1992) (KW1) Jugendkultur (KW2)	Die Dokumente thematisieren die Mediennutzung und Medienkompetenz von Jugendlichen. Im speziellen werden die Folgen von vermehrter Computernutzung und Auswirkung von Computerspielen beleuchtet.	Medienkompetenz Jugend Computerspiele

Table 1: Results for Simulated Information Needs and Queries from Keywords

The derived ground truth now can serve as baseline for the comparison of different retrieval tests configurations:

- Using only annotations (keywords)
- Using the query set together with annotations (keywords)
- Using only simulated queries

2.3 Pseudo test collections for learning to rank scientific articles

Berendsen et al. [2012a] automatically generate pseudo test collections to provide training material for learning to rank methods. Methods are proposed for generating pseudo test collections in the domain of digital libraries, where data are relatively sparse, but come with rich annotations. Their goal is to exploit document annotations, using the intuition that documents are annotated to make them better findable for heterogeneous information needs. Annotations and the associated documents are used as a source for queries and relevant documents.

With three different methods for sampling potentially relevant documents and two different methods for generating queries, a total of six pseudo test collection generation methods are proposed. Learning to rank (LTR) algorithms are trained on these generated collections and performance is compared to that of about ten well-known retrieval methods. In addition, Berendsen et al. [2012a] address the question when generalization is better: when training on a pseudo test collection or when training on an editorial collection. To answer this question, they make use of two editorial test collections: the CLEF 2007 and CLEF 2008 Domain Specific Track test collections. They find that training on the 2008 collection is better than training on

any of the generated pseudo test collections when tested on the 2007 topics. However, training on the 2007 topics is no better than training on several of the generated pseudo test collections, when evaluating on the 2008 topics.

It is also shown that a pseudo test collection can be useful for training even if it cannot be used for benchmarking in a Cranfield style evaluation campaign. To be more precise, even if a pseudo test collection does not rank ten different retrieval algorithms from the literature the same as editorial test collections do, it can still be successfully used to train an LTR model.

Note that this work is related to the work we reported on in deliverable 4.1 [Berendsen et al. 2011b], section 3.4.1., where we generate known-item queries from selected documents to create a pseudo test collection for known item search, which was then compared to a test collection created from purchase decisions obtained from transaction logs [Huurnink et al. 2010]. We extended this work with using annotations to group documents, generalizing to the ad hoc search task, where a query can have multiple relevant documents. We also used a different query generation mechanism. Finally, we pursued a different goal: rather than benchmarking retrieval systems, we train a learning to rank algorithm.

We are currently following up our latest work by developing a method to use knowledge derived from editorial judgments in the pseudo test collection generation pipeline. We are exploring the generation of ground truth in other domains, such as microblog search.

2.4 Simulated Relevance Assessments from Annotations

The exploitation of explicit annotations of documents for generating ground truth is tested in order to answer the research question, whether or not annotations can be used to simulate relevance assessments. Manual and automatic relevance assessments are compared in different variables using the bilingual GIRT collection (GESIS) which provides sets of queries, documents and manual relevance assessments for the purpose of comparison. Likewise, ad-hoc retrieval data from a previous CLEF test are available for further use.

The first step translates the given queries into controlled vocabularies such as keywords (i) and classifications (ii) using the MIND-server (GESIS). In addition, keywords will be automatically extracted (iii) from the manual assessed documents for the purpose of control. The generation of ground truth takes place via three different types of annotations (figure 2). Each of them will be used in two different tests. Since the gold standard of a manual relevance assessment in comparison with ad-hoc information retrieval data already exists (test A), the first examination (test B) will use the automatic relevance assessments as ground truth for measuring ad-hoc retrieval quality. A second examination (test C) will be a direct comparison of the manual and automatic baselines in identifying the overlap of relevant retrieved

documents. A further question is whether or not there will be a difference in recall (doc 5). In order to answer the research question, the threshold for an appropriate simulation of the gold standard has to be defined.

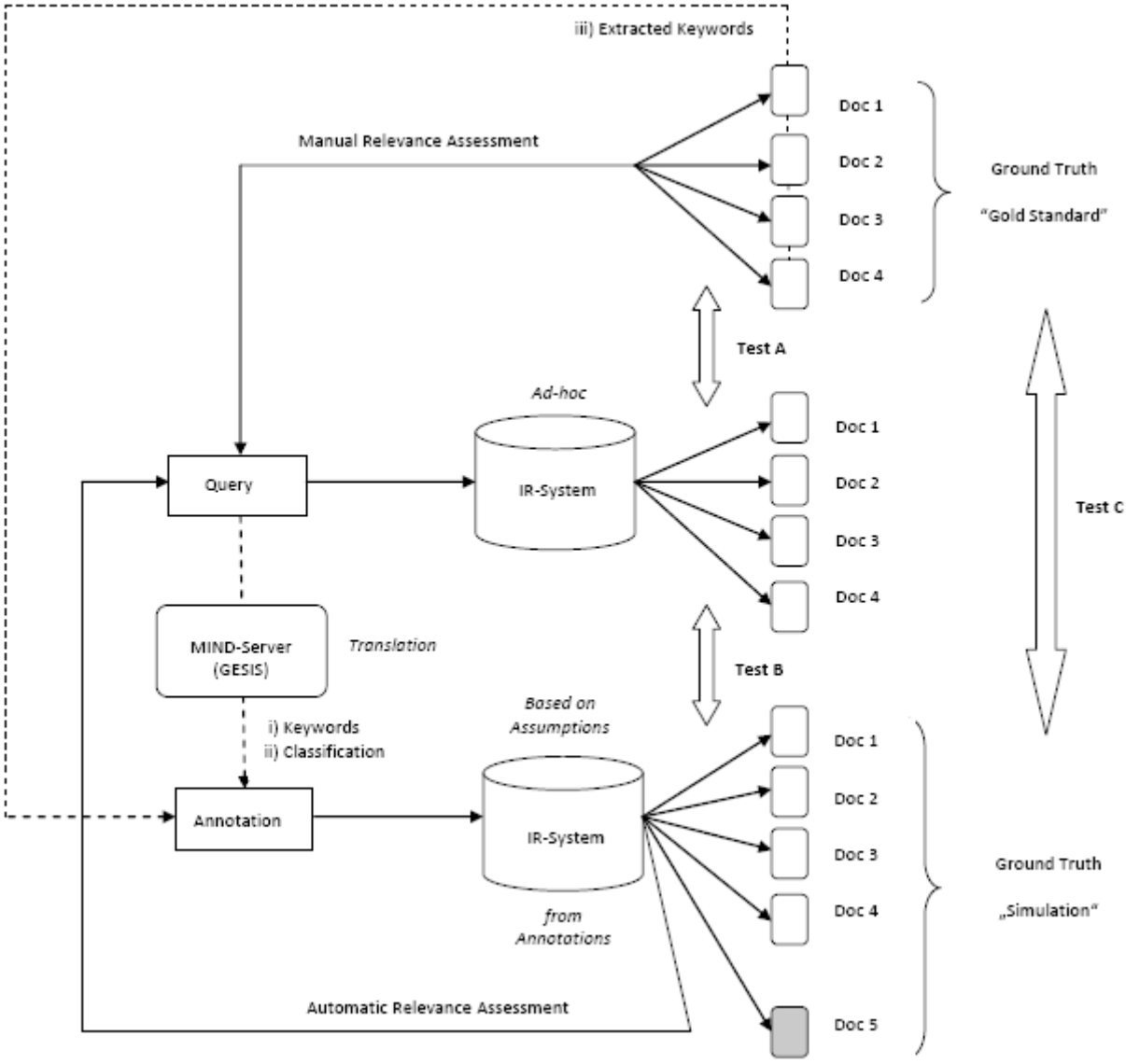

Figure 2: Overview Experiment II

2.5 Explicit annotations in Search for Innovation

The Search for Innovation use-case domain benefits, at least in some of its tasks, from the existence of numerous explicit annotations (metadata) in the patent corpus. Such annotations are either created together with the first version of a patent document (e.g. inventor, assignee, classification tags), or in subsequent processing

during the examination procedures (references to other documents, additional classification tags).

This metadata can, and has been used to create relevance judgements. For instance, classification tags have been used by Andersson [2010] as indicators of relevance between claims of patent applications, based on the prior observation of Krier and Zaccà [2002]. Patent classifications (the International Patent Classification - IPC, but even more so the European Classification - ECLA and the new Cooperative Patent Classification – CPC) are indeed very refined (e.g. CPC is estimated to have between 140k and 180k entries [EPO 2012]).

The classification tags do not however match the nature of the task at hand. A much better solution is to use the examination report and the list of relevant documents cited therein. We have shown before [Lupu et al. 2010] that such a method has a higher correlation to expert assessment than the pseudo-relevance method suggested by Soboroff [2001]. Even more, thanks to additional metadata in the patents (i.e. the priority numbers and, as their consequence, the family numbers), we can expand these relevance judgements using the method depicted in figure 3 below.

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

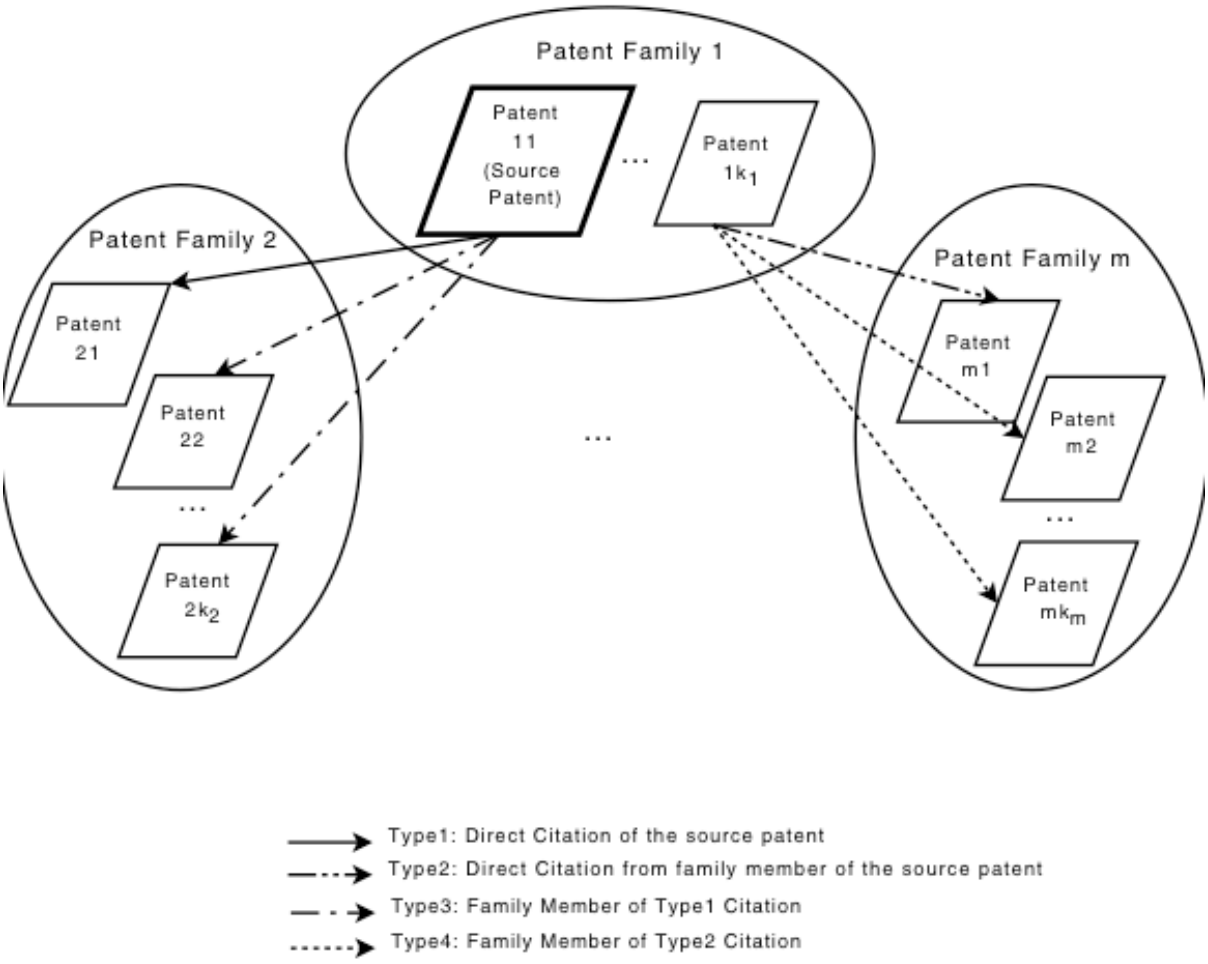

Figure 3: Patent Citation Extension used in CLEF-IP

More recently, we take the use of the examination report even further, and, for the time being, manually, extract paragraph relevance information for each citation. The evaluation results for this experiment are still pending the CLEF-IP 2012 campaign results, but as a consequence of this year's manual efforts, we have created a corpus to assist in the automatic extraction of such passage data.

3 Ground Truth from Log files

In deliverable 4.1 [Berendsen et al. 2011b], we reported at length on how analyzing usage data of search engines helps us to understand end user information needs. Also, we can mine such log files for ground truth to benchmark retrieval algorithms, or improve retrieval algorithms. In this section, we briefly comment on some recent

developments building on the work done in the finished Task 4.1 - Generating Ground Truth from Log files.

3.1 Result Disambiguation in Web People Search

As reported on in the previous deliverable [Berendsen et al. 2011b], we studied the log files of a people search engine to understand the types of queries end users pose, so that we can better develop and evaluate people search engines in accordance with end user information needs [Weerkamp et al. 2011, Berendsen et al. 2011a]. Weerkamp et al. [2011] give some recommendations for future work. One of them is to address the large ambiguity of many person names. This problem had already been studied extensively in the web search domain in the Web People Search (WePS) evaluation campaign [Artiles et al. 2007, 2009, 2010]. Person name queries were issued to a general purpose web search engine (Yahoo!) and participants were required to cluster search results such that each person being referred to in the search results would be associated with one cluster holding all documents referring to him or her. Berendsen et al. [2012b] found that state of the art algorithms for clustering people search results failed in this setting. The main cause: the many social media profiles in the search results of a dedicated people search engine. Social media profiles are often private, and even if not, their textual content is sparse. Therefore, a dedicated approach was needed. In the first step of a two-step strategy, non social media profiles and social media profiles were clustered separately, with distinct features. For social media profiles, some features were derived from transaction logs of the people search engine, following up on our research on mining ground truth from log files [Berendsen et al. 2011b]. These features, however, proved unsuccessful due to data sparsity. In the second step of the two stage strategy, both clusterings were merged, producing a final merged clustering which achieved state of the art performance compared with results obtained in the WePS campaign.

3.2 Interpreting clicks as relevance feedback

In contrast to Cranfield style benchmarking, companies with online search services often evaluate fewer systems (e.g. a current version and a version with a proposed new feature, see Kohavi et al. [2008] for a review), using more queries, and usage data of a vast amount of users.

One promising line of research is to interleave the result lists of two rankers and observe end user clicks on these interleaved lists to infer which of the two rankers has the better retrieval performance [Radlinski et al. 2008].

Hofmann et al. [2011] proposed a probabilistic method of interleaving ranked lists that allows to detect the better ranker in a pair of rankers with higher accuracy than

achieved before with other methods to construct interleaved rankings. However, a main limitation of the method of interleaving ranked lists is that it requires live clicks on the interleaved list each time two new rankers have to be compared. To overcome this problem, Hofmann et al. [2012] investigated if clicks on two rankers that were compared in the past can be reused to compare a new pair of rankers. They find that the probabilistic method of interleaving can indeed be used for this purpose but that it may be biased. To overcome this bias, they apply importance sampling and with this addition it is shown that historical comparison data can indeed be used to compare a new pair of rankers.

4 Alternative Retrieval Scenarios and Evaluation Metrics

This chapter provides some insights into research in alternative retrieval scenarios that are experimented with a CLEF2012 (in the PROMISE use case domain “Unlocking Culture” and in reputation management) as well as other applications at the Patolympics and in evaluating enterprise search solutions. Five different scenarios are introduced and described (some of which will be evaluated during the CLEF conference) all serving a common goal: to make information retrieval evaluation scenarios more realistic and user-centric.

4.1 CHIC2012 – Cultural Heritage in CLEF

The Cultural Heritage in CLEF (CHiC) pilot evaluation lab[1] aims at moving towards a systematic and large-scale evaluation of cultural heritage digital libraries and information access systems. The lab's goal is to increase our understanding on how to integrate examples from the cultural heritage community into a CLEF-style evaluation framework and how results can be fed back into the CH community. Data test collections and queries will come from the cultural heritage domain (in 2012 from Europeana) and tasks will contain a mix of conventional system-oriented evaluation scenarios (e.g. ad-hoc retrieval and semantic enrichment) for the CH domain, i.e. a variability task to present a particular good overview ("must sees") over the different objects types and categories in the collection targeted towards a casual user.

4.1.1 CHIC: Diversity Task

The diversity task requires systems to present a list of 12 objects (represents the first Europeana results page), which are relevant to the query and should present a particular good overview over the different object types and categories targeted towards a casual user, who might like the "best" documents possibly sorted into

[1] www.culturalheritageevaluation.org

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

"must sees" and "other possibilities." This task is about returning diverse objects and resembles the diversity tasks of the Interactive TREC track or the CLEF Image photo tracks. For CHIC, this task resembles a typical user of a cultural heritage information system, who would like to get an overview over what the system has with respect to a certain concept or what the best alternatives are. It is also a pilot task for this type of data collection. Documents returned should be as diverse as possible with respect to:

- media type of object (text, image, audio, video)
- content provider
- query category
- field match (which metadata field contains a query term)

We will test monolingual, bilingual and multilingual retrieval in 3 major European languages: English, French and German.

Topics: Topics are taken from real-life Europeana query topics and consist of a mixture of topical and named-entity queries. The 25 topics reflect real expressed user needs and are distributed according to query category statistics (mostly named entities, some topical queries etc.) but will be enhanced with suggested query categories that show different ambiguous aspects of a topic (e.g. topic = “Chardonne”, categories: person, place). More query categories can be suggested by participants.

Expected results: Participants are expected to submit 12 relevant results for all 25 topics in TREC-style format. More specifications on the result formatting will be released later.

Relevance assessments: Relevance assessments will be done manually by first collaboratively generating an assumed information need for the query and describing it (which will be used for later editions) and assessing the pooled documents for their relevance according to the query + information need + variability / diversity. If possible, we will compare 2 types of assessments: cultural heritage experts vs. “naive” users of cultural heritage information systems in order to be able to compare their assessments of relevance and variability.

Evaluation metrics: The evaluation metrics for the variabililty task will be the standard information retrieval measure of precision, particularly the standard measure mean average precision (MAP) and precision@k as well as diversity measures used in the Interactive TREC track like cluster-recall and intent-aware precision, which might be adapted to the diversity requirements set forth in this task.

4.1.2 CHIC: Semantic Enrichment Task

Task definition: The task requires systems to present a ranked list of at most 10 related concepts for a query to semantically enrich the query and / or guess the user’s information need or original query intent. Related concepts can be extracted

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

from Europeana data (internal information) or from other resources in the LOD cloud or other external resources (e.g. Wikipedia).

Europeana already enriches about 30% of its metadata objects with *concept names* and *place* (included in the test collection). It uses the following vocabularies for its included semantic enrichments, which can be explored further as well:

- GeoNames
- GEMET
- dbPedia

Semantic enrichment is an important task in information systems with short and therefore ambiguous queries like Europeana, which will support the information retrieval process either interactively (the user is asked for clarification, e.g. “Did you mean?”) or automatically (the query is automatically expanded with semantically related concepts to increase the likely search success). For CHIC, this task resembles a typical user interaction, where the system should react to an ambiguous query with a clarification request (or a result output as required in the variability task). We will offer the task and topics in 3 major European languages: English, French and German.

Additional Collections: For semantic enrichment, the Europeana Linked Open Data collections can also be used: Europeana released metadata on 2.5 million objects as linked open data in a pilot project. The data is represented in the Europeana Data Model (RDF) and encompasses collections from ca. 300 content providers. Other external resources are allowed but need to be specified in the description from participants. The objects described in the LOD dataset are included in the Europeana test collection, but the RDF format might be convenient for accessing object enrichments.

Topics: Topics are taken from real-life Europeana query topics and consist of a mixture of topical and named-entity queries. The 25 topics reflect real expressed user needs and are distributed according to query category statistics (mostly named entities, some topical queries etc.).

Expected results: Participants are expected to submit 10 ranked different terms or phrases for all 25 topics which express semantic enrichments for the query in the respective language and could be used for query expansion. More specifications on the result formatting will be released later.

Relevance assessments: Relevance will be assessed in 2 phases:

(1) First all submitted enrichments will be assessed manually for use in an interactive query expansion environment (e.g. “does this suggestion make sense with respect to the original query?).

(2) The submitted terms and phrases will be used in a query expansion experiment with a standard IR system, i.e. the enrichments will be individually

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

added to the query and submitted to the system. The results will be assessed according to ad-hoc retrieval standards.

Evaluation metrics: The evaluation metrics for the semantic enrichment task will be the standard information retrieval measure of precision (+precision@1 and @3) for the first phase of assessing just the submitted enrichments and the standard ad-hoc information retrieval measures for the second phase of assessing the submitted enrichments as query expansion variations.

A discussion of the tasks and the results can be found in the CHiC CLEF 2012 overview paper [Petras et al. 2012]

4.2 Application-Centric Black Box Evaluation / User Perception

In the following, we present a methodology that emphasizes application-centric evaluation. As laid out in this deliverable's introduction, the established Cranfield paradigm [Cleverdon 1997, Voorhees 2002] covers controlled experiments of information retrieval (IR) *systems* only, i.e. feeding a formulated query into a matching system and receiving a ranked list as output. Operational variables are eliminated in favor of laboratory-like conditions. Since one of the goals of PROMISE is stepping out of the laboratory into operational environments, operational variables are indeed relevant. Within such an operational setting, IR applications are used as supporting tools for knowledge-intensive business processes rather than being scrutinized in isolation as in academia. The intention is thus to evaluate applications as a whole as they are employed in the industry, where more components factor into actual performance than only query-document matching.

An *application* in this context is thought of as the combination of an IR system, the covered data, the application interface and the overall application configuration, as in the following figure:

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

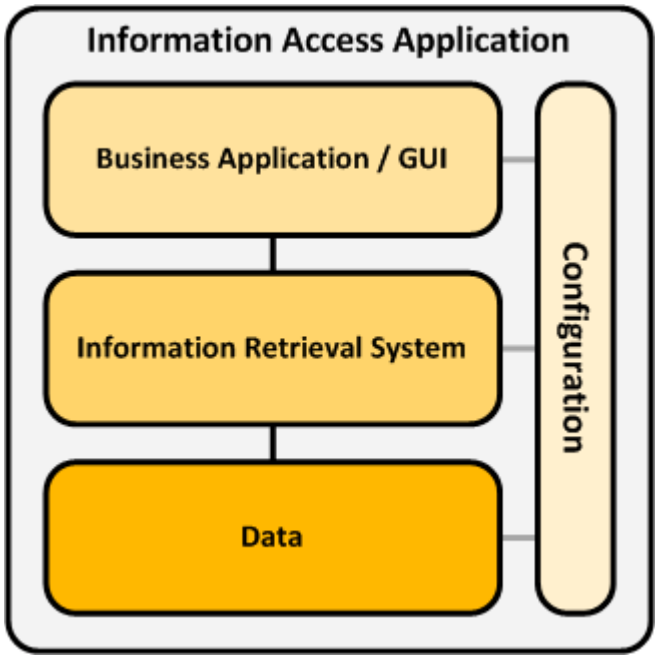

Figure 4: Information Access Application

The methodology as presented treats applications as black boxes where only publicly accessible interfaces can be used. This approach was taken because the methodology relies on coarse measurements and it is unclear (on-going research) if and how much a glass box approach would benefit the evaluation, especially in the context of a generic methodology.

The methodology is based on a hierarchical tree of criteria. Depending on the use case domain of the applications to be evaluated, the tree is pruned to produce an applicable subset of criteria. These criteria are assessed by many simple tests requiring only basic public access to a search application. By performing coarse, orthogonal tests and then aggregating them, we assess the application as a whole (more on this in previous PROMISE deliverables D4.1 [Berendsen et al. 2011b] and D2.2 [Järvelin et al. 2012]. Additionally, the large number of tests is required to cover all application components as previously described.

The test scripts omit the actual users by design. The processes in the context of an IR application are very knowledge-intensive and users have a lot of implicit knowledge. Therefore, the users' needs and use cases are implicitly modeled in the tests. This is possible by focusing on use case domains and modeling "prototypical users". This way, the testers involved in the evaluation need not be experts in the use case domain of each application and evaluation costs can be kept low.

Criteria and tests are based on application features and behaviors that are beneficial to a user's search experience. Low scores on tests indicate an aspect of an application where a user's search experience and performance is impaired. The

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

chosen measure for the evaluation is therefore linked to an *estimate of user perception* of an application.

Here is an example of a criterion and test description:

Index Completeness
Assumption
Users expect to potentially find all documents that can be publicly accessed in any way on the site (namely, through browsing the site) when using the search facility.
Irregularity
Publicly accessible documents (known through browsing or obtaining a direct link) cannot be found using the search facility.
Root causes
The index is incomplete – documents/sets of documents are missing The index is incomplete – the index is out of date (→ Freshness) The index is incomplete – documents of certain types are missing (→ Format support)
Tests
The content scope of the test is decidedly narrow. No linked resources in the application are expected to be accessible through the search. E.g. if a retailer owns another shopping outlet, the latter's products need not be found. 1. Tester locates three documents that match the following criteria: a. at least 5 clicks to locate document b. document is at least 3 levels from root (as determined by URL) (optional if URL rewriting is used in application) c. URL is at least 100 characters long (optional as above) 2. If no documents matching criteria 1) are found → abort 3. Tester extracts a characteristic phrase (this should be defined in a central location of the doc) from the document 4. Tester searches for the document 5. Score: number of documents that can be located in the top 10 search results (0, 1, 2, 3)

Table 2: Criterion and Test Description

As a validation inside PROMISE, a campaign was conducted where each participating PROMISE partner was asked to identify ten target sites, which they would evaluate. The sites were required to be based in the partners' country, would fit in PROMISE use case domains and/or belong to well-known or economically strong organizations (implicit "enterprise search" use case). Partners were provided with the test scripts and an accompanying scoring sheet. The final evaluation of the results is a currently on-going effort.

4.3 Expert finding and expert profiling

It has been recognized that expert finding is a very important retrieval task in enterprise environments. Expert profiling is a complementary task, where the query is an expert, and the desired result is a profile consisting of expertise areas. [Balog 2008]. Balog et al. [2007] released a test collection with a new kind of ground truth for both tasks: experts from the University of Tilburg had selected expertise areas from a knowledge base for their own profile. For expert finding, then, for each expertise area the ground truth is created by all experts who included this area in their profile. For expert profiling, the ground truth is created by the expert the end user wants to profile. Motivated by the sparseness of the judgments: (each area is only selected by very few experts) we are currently finishing work on an experiment where we asked experts to judge profiles that were generated for them by a combination of state of the art algorithms. This way, we expect to be able to more reliably rank retrieval algorithms for expert profiling. We will be releasing a new test collection as an outcome of this assessment experiment. We also gave experts the opportunity to give free text feedback on their generated profiles. By doing a content analysis of these comments we find aspects that are of importance for developing and evaluating expert profiling systems. One of the aspects experts voiced was that redundancy in profiles occurred but is unwanted. De Rijke et al. [2010] proposed a new set of metrics to address this issue. For a recent overview on expertise retrieval, see Balog et al. [2012].

4.4 Reputation Management

At CLEF 2012, one of the labs to be organized is RepLab, dealing with reputation management: “While traditional reputation analysis was based mostly on manual analysis (clipping from media, surveys, etc.), the key value from online media comes from the ability of processing, understanding and aggregating potentially huge streams of facts and opinions about a company or individual”[2]. Information to be mined includes answers to questions such as: What is the general state of opinion about a company/individual in online media? What are its perceived strengths and weaknesses as compared to its peers/competitors? How is the company positioned with respect to its strategic market? Can incoming threats to its reputation be detected early enough to be neutralized before they effectively affect reputation? In this context, Natural Language Processing plays a key, enabling role and we are already witnessing an unprecedented demand for text mining software in this area. Note that while the area of opinion mining has made significant advances in the last few years, most tangible progress has been focused on products. However, mining and understanding opinions about companies and individuals is, in general, a much

[2] http://www.limosine-project.eu/events/replab2012.

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

harder and less understood problem. The aim of this lab is to bring together the Information Access research community with representatives from the Online Reputation Management industry, with the goals of (i) establishing a five-year roadmap that includes a description of the language technologies required in terms of resources, algorithms, and applications; (ii) specifying suitable evaluation methodologies and metrics; and (iii) developing of test collections that enable systematic comparison of algorithms and reliable benchmarking of commercial systems." For recent developments in this area, see Spina et al. [2012a, 2012b].

4.5 PatOlympics 2012

The evaluation efforts of CLEF-IP were complemented with an interactive evaluation session called PatOlympics [Lupu 2011]. The experiment brought together examiners and IR scientists and, while keeping as much as possible from the reliability and fairness of the benchmarking-style of tests, it generated relevance judgements on the fly, based on the input of the patent experts. The first two PatOlympics focused on the quality of the retrieval, with scores publicly displayed and dynamically updated for all participants. The latest instance of the event dropped the constraints of previous versions in order to focus on the user interaction. In this sense, we collaborated with the new PROMISE partner in Denmark (RSLIS) and recorded the interactions between users (experts) and systems.

5 Future Work and Conclusions

This report described ongoing research in the area of generating ground truth from annotations and collections and alternative evaluation scenarios and metrics. While the first task is research into making evaluation more efficient by saving time- and resource-consuming manual labor for mostly relevance assessments, the intention of alternative evaluation scenarios is commonly to find more realistic evaluation tests than the Cranfield paradigm. The report presented four experiments within the PROMISE project on generating automatic ground truth and five different alternative evaluation scenarios. It is too early to judge their success in the wider field, but the application of several of the proposed scenarios within the active CLEF laboratory environment provides hopes for further use.

Several projects within the PROMISE environment dealing with alternative evaluation methodologies are still ongoing. All studies have in common that they argue to broaden the scope of evaluation research including user-centric factors. Based on a use-case motivated evaluation approach [Karlgren et al. 2011] a simulated search session experiment investigating the effect of session length, query reformulation and results viewed has been conducted. The results will be published in future and inform follow up studies in this area.

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

References

[Ali & Sufyan Beg 2011] R. Ali, M. M. Sufyan Beg, “An overview of web search evaluation methods”, in ‘Computers and Electrical Engineering”, 37, pp. 835-838, 2011.

[Amitay et al.2004] E. Amitay, D. Carmel, R. Lempel, and A. Soffer, “Scaling IR-system evaluation using term relevance sets,” in ‘Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval’, pp. 10–17, New York, NY, USA, 2004.

[Andersson 2010] L. Andersson, “A vector space analysis of Swedish patent claims with different linguistic indices”, in ‘Proc. of PaIR’, 2010.

[Artiles et al. 2007] J. Artiles, J. Gonzalo, 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’, pp. 64–69, 2007.

[Artiles et al. 2009] J. Artiles, J. Gonzalo, 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. 2010] J. Artiles, A. Borthwick, J. Gonzalo, S. Sekine, E. Amigo, “WePS-3 evaluation campaign: Overview of the web people search clustering and attribute extraction tasks”, in ‘CLEF 2010 Working Notes’, 2010.

[Asadi et al. 2011] N. Asadi, D. Metzler, T. Elsayed, J. Lin, “Pseudo test collections for learning web search ranking functions”, in ‘SIGIR ’11’, pp. 1073–1082, ACM, 2011.

[Balog 2008] K. Balog, "People search in the enterprise", in ‘SIGIR Forum’, vol. 42, no. 2, pp. 103, December 2008.

[Balog et al. 2007] K. Balog, T. Bogers, L.A. Azzopardi, M. de Rijke, 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’, New York, NY, USA, ACM Press, pp. 551-558, 2007.

[Balog et al. 2012] K. Balog, Y. Fang, M. de Rijke, P. Serdyukov, L. Si, "Expertise retrieval", in ‘Foundations and Trends in Information Retrieval’, vol. 6, no. 2-3, pp. 127-256, August, 2012.

[Bao et al. 2007] S. Bao, G. Xue, X. Wu, Y. Yu, B. Fei, Z. Su, “Optimizing web search using social annotations”, in ‘WWW ’07 Proceedings of the 16th International Conference on World Wide Web, pp. 501–510, New York, NY, USA, 2007.

[Beaulieu 1997] M. Beaulieu, “Experiments with interfaces to support query expansion”, in ‘Journal of Documentation’, 53 (1), pp 8-19, 1997.

[Beitzel et al. 2003b] S. M. Beitzel, E. C. Jensen, A. Chowdhury, D. Grossman, “Using titles and category names from editor-driven taxonomies for automatic evaluation”, in ‘CIKM '03 Proceedings of the twelfth International Conference on Information and Knowledge Management’, pp. 17–23, New Orleans, LA, USA 2003.

[Beitzel et al. 2003c] S. M. Beitzel, E. C. Jensen, A. Chowdhury, D. Grossman, 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', pp. 373–374, New York, NY, USA, 2003.

[Beitzel et al.2003a] S. M. Beitzel, E. C. Jensen, A. Chowdhury, D. Grossman, "Semi-automatic evaluation via editor-driven taxonomies", in 'Terabyte Collections Workshop at SIGIR'03, New York, NY, USA, 2003.

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

[Berendsen et al. 2011b] R. Berendsen, G.M. Di Nunzio, M. Gäde, J. Karlgren, M. Lupu, S. Rietberger, J. Stiller, "First report on alternative evaluation methodology. PROMISE Deliverable 4.1", PROMISE project, 2011.

[Berendsen et al. 2012a] R. Berendsen, E. Tsagkias, M. de Rijke, E. Meij, "Generating pseudo test collections for learning to rank scientific articles", in 'CLEF 2012: Conference and Labs of the Evaluation Forum', Rome, Italy, Springer, September 2012.

[Berendsen et al. 2012b] R. Berendsen, B. Kovachev, E. Nastou, M. de Rijke, W. Weerkamp, "Result Disambiguation in Web People Search", in 'ECIR 2012: 34th European Conference on Information Retrieval', pp. 146-157, Barcelona, April 2012.

[Bosca & Dini 2009] A. Bosca, L. Dini, "Evaluating systems for multilingual and multimodal information access lecture notes", in 'Computer Science', 5706, pp. 42-49, DOI: 10.1007/978-3-642-04447-2_4, 2009.

[Bruza et al. 2000] P. Bruza, R. MacArthur, S. Dennis, "Interactive internet search: Keyword, directory and query reformulation mechanisms compared", in 'Proceedings of ACM SIGIR '00', 2000.

[Cecchini et al. 2011] R. L. Cecchini, C. M. Lorenzetti, A. G. Maguitman, F. Menczer, "A semantic framework for evaluating topical search methods", in 'CLEIej', 14 (1), paper 2, http://www.clei.cl/cleiej/paper.php?id=211, 2011.

[Chowdhury & Soboroff 2002] A. Chowdhury, I. Soboroff, "Automatic evaluation of world wide web search services", 'in 'Proceedings of SIGIR'02', pp. 421 - 422, New York, NY, USA, 2002.

[Cleverdon 1997] C. Cleverdon, "The Cranfield tests on index language devices" in K. S. Jones & P. Willett (Eds.), 'Readings in information retrieval', pp. 47–59, San Francisco, 1997.

[De Rijke et al. 2010] M. de Rijke, K. Balog, T. Bogers, A. van den Bosch, "On the evaluation of entity profiles", in 'CLEF 2010: Conference on Multilingual and Multimodal Information Access Evaluation', Padova, September 2010.

[Dolin et al. 1998] R. Dolin, D. Agrawal, A. E. Abbadi, J. Pearlman, "Using automated classification for summarizing and selecting heterogeneous information sources", in 'D-Lib Magazine' January 1998.

[EPO 2012] European Patent Office, CPC Workshop for External Users, Main presentation, Vienna 23-03-2012, http://www.cooperativepatentclassification.org/publications/WorkshopMarchVienna.pdf, 2012.

[French et al. 2001] J.C. French, A.L. Powell, F. Gey, N. Perelman, “Exploiting a controlled vocabulary to improve collection selection and retrieval effectiveness”, in ‘10th International Conference on Information and Knowledge Management’, pp. 199–206, 2001.

[Gey & Jiang 2000] F. C. Gey, H. Jiang, “English-German cross-language retrieval for the GIRT collection - exploiting a multilingual thesaurus”, in ‘Proceedings of the Eighth Text Retrieval Conference, TREC8’, 2000.

[Harmandas et al. 1997] V. Harmandas, M. Sanderson, M. D. Dunlop, “Image retrieval by hypertext links,” in 'Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval‘, pp. 296–303, New York, NY, USA, 1997.

[Harter & Cheng 1996] S. P. Harter, Y.-R. Cheng, “Colinked descriptors: Improving vocabulary selection for end-user searching”, in ‘Journal of the American Society for Information Science’, 47, pp. 311-325, 1996.

[Haveliwala et al. 2002] T. H. Haveliwala, A. Gionis, D. Klein, P. Indyk, “Evaluating strategies for similarity search on the web,” in ‘Proceedings of the 11th International Conference on World Wide Web‘, pp. 432–442, New York, NY, USA, 2002.

[Hawking et al. 2004] D. Hawking, F. Crimmins, N. Craswell, “How valuable is external link evidence when searching enterprise webs?,” in ‘Proceedings of the 15th Australasian Database Conference, vol. 27, pp. 77–84, Darlinghurst, Australia, 2004.

[He et al. 2011a] J. He, M. de Rijke, M. Sevenster, R. van Ommering, Y. Qian, "Generating links to background knowledge: A case study using narrative radiology reports", in ‘20th ACM Conference on Information and Knowledge Management (CIKM 2011)’, pp. 1867--1876, Glasgow, ACM, October 2011.

[He et al. 2011b] J. He, M. de Rijke, M. Sevenster, "Generating links to background knowledge for medical content", in ‘Second International Workshop on Web Science and Information Exchange in the Medical Web (MedEX 2011)’, Glasgow, ACM, October, 2011.

[Hersh & Kim 2006] W. Hersh, E. Kim, “The impact of relevance judgments and data fusion on results of image retrieval test collections”, in ‘Proceedings of the 2nd MUSCLE/ImageCLEF Workshop on Image and Video Retrieval Evaluation’,p. 29–38, Alicante, Spain, 2006.

[Hersh et al. 1996] W. R. Hersh, J. Pentecost, D. Hickam, “A task-oriented approach to information retrieval evaluation”, in ‘Journal of American Society for Information Science’, 47, pp. 50-56, 1996.

[Hersh et al. 2000] W. R. Hersh, S. Price, L. Donohoe, “Assessing thesaurus-based query expansion using the UMLS metathesaurus”, in ‘Proceedings of AMIA Annual Fall Symposium”, pp. 344–348, 2000.

[Hofmann et al. 2011] K. Hofmann, S. Whiteson, M. de Rijke, "A probabilistic method for inferring preferences from clicks", in ‘20th ACM Conference on Information and Knowledge Management (CIKM 2011)’, pp. 249-258, Glasgow, ACM, October 2011.

[Hofmann et al. 2012] K. Hofmann, S. Whiteson, M. de Rijke, "Estimating interleaved comparison outcomes from historical click data", in ‘CIKM 2012: 21st ACM Conference on Information and Knowledge Management’, ACM, October 2012.

[Hotho et al. 2006] A. Hotho, R. Jaschke, C. Schmitz, G. Stumme, “Information retrieval in folksonomies: Search and ranking”, in: ‘Proceedings of ESWC ’06’, pp.411-426, 2006.

[Humphrey 1999] S. M. Humphrey, “Automatic indexing of documents from journal descriptors: A preliminary investigation”, in ‘Journal of the American Society for Information Science, 50 (8), pp. 661-674, 1999.

[Huumink et al. 2010] B. Huurnink, K. Hofmann, M. de Rijke, M. Bron, "Validating query simulators: An experiment using commercial searches and purchases", in ‘CLEF 2010: Conference on Multilingual and Multimodal Information Access Evaluation’, Padova, Springer, September 2010.

[Iivonen & Sonnenwald 1998] M. Iivonen, D. H. Sonnenwald, “From translation to navigation of different discourses: A model of search term selection during the pre-online stage of the search process”, in ‘JASIS’, 49 (4), pp. 312-326, 1998.

[Ingwersen & Järvelin 2005] P. Ingwersen, K. Järvelin. “The turn: Integration of information seeking and retrieval in context. The Kluwer International Series on Information Retrieval, Springer, 2005.

[Järvelin et al. 2012] A. Järvelin, G. Eriksson, P. Hansen, T. Tsikrika, A. Garcia Seco de Herrera, M. Lupu, M. Gäde, V. Petras, S. Rietberger, M. Braschler, R. Berendsen, “Revised Specification of Evaluation Tasks. PROMISE Deliverable 2.2”, PROMISE project, 2012.

[Jensen et al. 2007] E. C. Jensen, S. M. Beitzel, A. Chowdhury, O. Frieder, “Repeatable evaluation of search services in dynamic environments,” in, 'ACM Transactions on Information Systems (TOIS)’, vol. 26, no. 1, p. 1, doi:10.1145/ 1292591.1292592, 2007.

[Jin et al. 2002] R. Jin, L. Si, A. G. Hauptman, J. Callan, “Language model for IR using collection information”, in ‘Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval’, pp. 419–420, New York NY, USA, 2002.

[Joho et al. 2004] H. Joho, M. Sanderson, M. Beaulieu, “A study of user interaction with a concept-based interactive query expansion support tool”, in ‘Proceedings of the 26th European Conference in Information Retrieval’, pp. 42–56, 2004.

[Kamps 2004] J. Kamps, “Improving retrieval effectiveness by re-ranking documents based on controlled vocabulary”, in ‘The 21th European Conference on Information Retrieval’, 2004.

[Karlgren et al. 2011] J. Karlgren, A. Järvelin, G. Eriksson, P. Hansen, „Use cases as components of information access evaluation”, in ‘Proc. of DESIRE’, 2011.

[Kohavi et al. 2008] R. Kohavi, R. Longbotham, D. Sommerfield, R. M. Henne, “Controlled experiments on the web: Survey and practical guide”, in ‘Data Mining and Knowledge Discovery’, vol. 18, No. 1., pp. 140-181, February 2009.

[Krier & Zaccà 2002] M. Krier, F. Zaccà, “Automatic categorization applications at the European patent office”, in ‘World Patent Information’, vol. 24(3), pp. 187-196, 2002.

[Liu et al. 2009] H. Liu, R. Song, J.-Y. Nie, J.-R. Wen, “Building a test collection for evaluating search result diversity: A preliminary study”, in ‘Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation’, pp. 31–32, 2009.

[Lupu 2011] M. Lupu, "PatOlympics: An infrastructure for interactive evaluation of patent retrieval tools" in 'Proc. of DESIRE', 2011.

[Lupu et al. 2010] M. Lupu, F. Piroi, A. Hanbury, "Aspects and analysis of patent test collections", in 'Proc. of PaIR', 2010.

[Maguitman et al. 2010] A. G. Maguitman, R. L. Cecchini, C. M. Lorenzetti, F. Menczer, "Using topic ontologies and semantic similarity data to evaluate topical search", in 'Proceedings of the 36th Latin American Informatics Conference (CLEI)', 2010.

[Markines et al. 2009] B. Markines, C. Cattuto, F. Menczer, D. Benz, A. Hotho, G. Stumme, "Evaluating similarity measures for emergent semantics of social tagging", in 'Proceedings of the 18th International World Wide Web Conference, pp. 641-650, 2009.

[Menczer 2003] F. Menczer, "Semi-supervised evaluation of search engines via semantic mapping", submitted to 'WWW '03', http://dollar.biz.uiowa.edu/~fil/Papers/engines.pdf., 2003.

[Morrison 2008] J. P. Morrison, "Tagging and searching: Search retrieval effectiveness folksonomies on the world wide web", in 'Information Processing & Management' 44 (4), pp. 1562-1579, 2008.

[Noll & Meinel 2007] M. G. Noll, C. Meinel, "Web search personalization via social bookmarking and tagging, in 'Proceedings of ISWC '07', 2007.

[Petras et al. 2012] V. Petras, N. Ferro, M. Gäde, A. Isaac, M. Kleineberg, I. Masiero, M. Nicchio, J. Stiller, „Cultural Heritage in CLEF (CHiC) Overview 2012", in "CLEF 2012 Working Notes". Rome, Italy, September 17-20 2012.

[Petras 2005] V. Petras, "GIRT and the use of subject metadata for retrieval", in 'Multilingual Information Access for Text, Speech and Images: Results of the Fifth CLEF Evaluation Campaign, CLEF '04', pp. 219–225, Berlin, 2005.

[Petras et al. 2002] V. Petras, N. Perelman, F. Gey, "Using thesauri in cross-language retrieval of German and French indexed collections", in 'Proceedings of the CLEF '02 Workshop', 2002.

[Radlinski et al. 2008] F. Radlinski, M. Kurup, T. Joachims, "How does clickthrough data reflect retrieval quality?" in J. G. Shanahan, S. Amer-Yahia, I. Manolescu, Y. Zhang, D. A. Evans, A. Kolcz, K.-S. Choi, A. Chowdhury (Eds.), 'CIKM', pp. 43 – 52. ACM, 2008.

[Ramage et al. 2009] D. Ramage, P. Heymann, C. D. Manning, H. Garcia-Molina, "Clustering the tagged web", in 'Proceedings of WSDM', 2009.

[Sanderson 2010] M. Sanderson, "Test collection based evaluation of information retrieval systems", in 'Foundations and Trends in Information Retrieval', 4, pp.247–375, 2010.

[Shiri et al. 2002] A. A. Shiri, C. Revie, G. Chowdhury, "Thesaurus-enhanced search interfaces", in 'Journal of Information Science', 28 (2), pp. 111–122, 2002.

[Soboroff et al. 2001] I. Soboroff, C. Nicholas, P. Cahan, "Ranking retrieval systems without relevance judgments," in 'Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval', pp. 66–73, New Orleans, Louisiana, United States, DOI:10.1145/383952.383961, 2001.

[Soergel 1997] D. Soergel, “Multilingual thesauri in cross-language text and speech retrieval”, in ‘AAAI Symposium on Cross-Language Text and Speech Retrieval’, 1997.

[Spina et al. 2012a] D. Spina, E. Meij, M. de Rijke, A. Oghina, B.M. Thuong, M. Breuss, "Identifying entity aspects in microblog posts", in ‘SIGIR '12: 35th international ACM SIGIR conference on Research and development in information retrieval’, Portland, Oregon, ACM, August, 2012.

[Spina et al. 2012b] D. Spina, E. Meij, A. Oghina, M. Bui, M. Breuss, M. de Rijke, "A corpus for entity profiling in microblog posts", in ‘REC 2012 Workshop on Language Engineering for Online Reputation Management’, Istanbul, May, 2012.

[Suomela & Kekäläinen 2005] S. Suomela, J. Kekäläinen, “Ontology as a search-tool: A study of real users' query formulation with and without conceptual support”, in ‘Proceedings of ECIR ‘05’, pp. 315-329, Berlin, Heidelberg, 2005.

[Taghva et al. 1999] K. Taghva, J. Borsack, A. Condit, "The effectiveness of thesauri-aided retrieval", in ‘Proceedings of IS&T/SPIE ‘99 International Symposium on Electronic Imaging Science and Technology’, pp. 202-211, 1999.

[Trant 2009] J. Trant, “Studying social tagging and folksonomies: a review and framework”, in ’The Journal of Digital Information, 10 (1), http://dlist.sir.arizona.edu/2595/, 2009.

[Vallet et al. 2010] D. Vallet, I. Cantador, J. M. Jose, “Personalizing web search with folksonomy-based user and document profiles”, in ‘Proceedings of ECIR ‘10’, pp. 420-431, 2010.

[Voorhees 2002] E. M. Voorhees. “The philosophy of information retrieval evaluation”, in ‘Evaluation of Cross-Language Information Retrieval Systems. Proceedings of CLEF 2001’, 2406, Lecture Notes in Computer Science, pp. 355–370, 2002.

[Weerkamp et al. 2011] W. Weerkamp, B. Kovachev, R. Berendsen, E. , K. Balog, M. de Rijke, "People searching for people: Analysis of a people search engine log", in ‘34th Annual International ACM SIGIR Conference (SIGIR 2011)’, Beijing, ACM, pp. 45-54, July, 2011.

[Xu et al. 2008] S. Xu, S. Bao, B. Fei, Z. Su, Y. Yu, “Exploring folksonomy for personalized search,” in 'Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval‘, pp. 155–162, New York, NY, USA, 2008.

[Zhou et al. 2008] D. Zhou, J. Bian, S. Zheng, H. Zha, C.L: Giles, “Exploring social annotations for information retrieval”, in ‚Proceedings of the 17th International Conference on World Wide Web‘, pp. 715–724, New York, NY, USA, 2008.