CLEF, CLEF 2010, and PROMISEs: Perspectives for the Cross-Language Evaluation Forum

Nicola Ferro
Department of Information Engineering
University of Padova
Via Gradenigo, 6/b – 35131 Padova, Italy
ferro@dei.unipd.it

ABSTRACT

After ten years of increasingly successful evaluation campaigns, the *Cross-Language Evaluation Forum (CLEF)* has come to an appropriate moment to assess what has been achieved in this decade and also to consider future directions and how to renew and complement it. This paper will provide a brief summary of the most significant results achieved by CLEF in the past ten years, it will describe the new format and organization for CLEF which is being experimented for the first time in CLEF 2010, and it will discuss some future perspective for CLEF, beyond 2010.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Information filtering, Relevance feedback, Retrieval models, Search process*; H.3.7 [**Information Storage and Retrieval**]: Digital Libraries—*Systems issues, User issues*; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*Graphical user interfaces*

General Terms

Algorithms, Design, Experimentation

1. INTRODUCTION

The growth of the Internet has been exponential with respect to the number of users and languages used regularly for global information dissemination. With the advance of broadband access and the evolution of both wired and wireless connection modes, users are now not only information consumers, but also information producers: they create their own content, augment existing material through annotations (e.g. adding tags and comments) and links, mix and mash up different media and applications within a dynamic and collaborative information space. The expectations and habits of users are constantly changing, together with the ways in which they interact with content and services, often creating new and original ways of exploiting them.

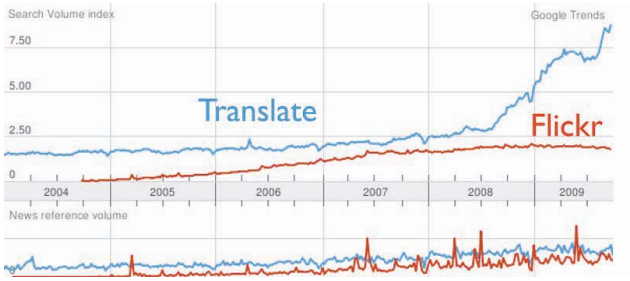

Figure 1: Google search trends for "translate" (in blue) and "flickr" (in red).

In this evolving scenario, language and media barriers are no longer seen as inviolable and they are constantly crossed and mixed to provide content that can be accessed on a global scale within a multicultural and multilingual setting.

Therefore, users need to be able to co-operate and communicate in a way which crosses language and media boundaries and goes beyond separate search in diverse media/languages, but which exploits the interactions between different languages and media. Indeed, language and media barriers are no more perceived simply as an "obstacle" to the retrieval of relevant information resources, but also represent a challenge for the whole communication process, i.e. information access and exchange.

As an indicator of these emerging user needs, let us consider the Google search trends[1] for two distinct queries – "translate" and "flickr" – as well as the volume of discussion they raised in the news. Figure 1 shows how much the interest of Internet users on these two topics has been growing constantly over the years and their needs are becoming more compelling: the word "translate" has tripled its relative frequency in Google searches over the last 15 months while the word "flickr" has almost doubled.

It is important to understand that the trends for "translate" and "flickr" queries are not two separate and independent phenomena, rather they are starting to interact with one another. This clearly emerges in Figure 2, where the search trends for the query "translate flickr" are shown: towards the end of 2008, users began to look for ways to access Flickr contents (images) in a multilingual way (text), indicating the need for information systems able to cross and

[1] `http://www.google.com/trends`
Google trends computes how many searches have been done for the entered terms, relative to the total number of searches done on Google over time.

Figure 2: Google search trends for "translate flickr".

mix language and media barriers.

In addition, it is well-known that coupling full text search with visual image search can improve information access effectiveness in a multilingual context [7]. Users can estimate the relevance of the items found from the images even if they are not familiar with the language used for the textual part, and images themselves are totally language independent.

This need for mixed multimedia and multilingual information access is not limited to the Web but concerns other important domains, such as patents. As you can note from Figure 3, patents can be filed in multiple languages and can contain multimedia elements, such as images or technical diagrams, that further explain the contents of the patent. Patent Search is an essentially multi-lingual problem. To be valid and legally defensible, the filed patent must be shown to be the first publication of the invention described in the patent. Publication in this case means any generally available description of the patent regardless of the language in which the invention is described. Thus, for example, a patent filed in English to the European Patent Office can be shown to be invalid by an academic paper in Russian [11].

As a consequence of this challenging scenario, information systems are becoming increasingly complex: they need to satisfy user needs and carry out tasks that are becoming progressively more complicated and cross language and media barriers; moreover, they have to manage increasing amounts of information which is often heterogeneous and demands for insightful access to it. Therefore, their design and development requires the integration of components and technologies coming from different areas and domains, which are rarely present in a single research group, as well

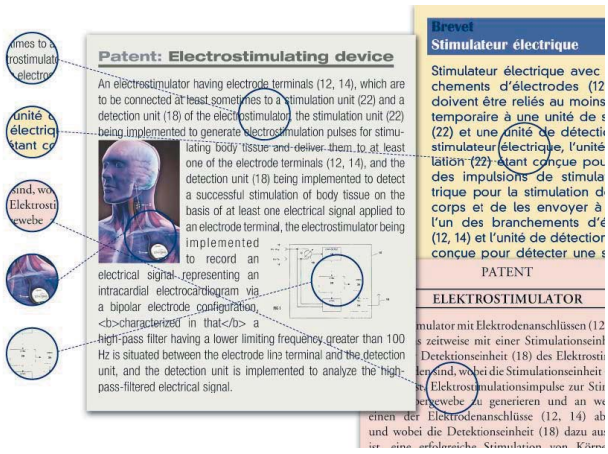

Figure 3: Example of multilingual and multimedia patent.

as the gathering of researchers and developers with multidisciplinary competencies able not only to go into the details of their own specific sector/component, but also to obtain an overall comprehension of the big picture and the interactions with the other domains. Nevertheless, complexity is not only intrinsic to the information systems themselves, but it also concerns the context in which these systems operate. Indeed, if we are to continue advancing the state-of-the-art in information access technologies, we need to understand a new breed of users who are performing different kinds of tasks within varying domains, often acting within communities to find and produce information not only for themselves, but also to share with other users. To this end, we must study and evaluate the interaction among four main entities: users, their tasks, languages, and multimedia content to help understand how these factors impact on the design and development of multilingual and multimedia information systems [4].

Therefore, we consider experimental evaluation – both laboratory and interactive – a key means for supporting and fostering the development of multilingual and multimedia information systems which are more adherent to the new user needs in order to ensure that they meet the expected user requirements, provide the desired effectiveness and efficiency, guarantee the required robustness and reliability, and operate with the necessary scalability.

The paper is organized as follows: Section 2 discusses the achievements of the first ten years of CLEF, which provide us with the indispensable basis for being able to start tackling the challenges described above. Section 3 present how we are renewing and evolving CLEF to start moving in the envisioned direction. Finally, Section 4 introduces our future plans, beyond CLEF 2010, towards "next generation"Êevaluation campaigns.

2. THE FIRST TEN YEARS OF CLEF

The *Cross-Language Evaluation Forum (CLEF)* has been launched as an European initiative in early 2000 with the following objectives: "to develop and maintain an infrastructure for the testing and evaluation of information retrieval systems operating on European languages, in both monolingual and cross-language contexts, and to create test-suites of reusable data that can be employed by system developers for benchmarking purposes" [8]. Although it is true to say that this basic idea remains at the core of our activity, over the years our range of interest and our interpretation of these initial objectives have both widened and deepened [5].

Indeed, in 2000 the CLEF focus was on text and document retrieval but, over the years, different kinds of text retrieval across languages (e.g. question answering and geographic *Information Retrieval (IR)* as well) and different kinds of media (e.g. images and speech) have been investigated. The goal has been not only to meet but also to anticipate the emerging needs of the R&D community and to encourage the development of next generation multilingual and multimedia IR systems. This has meant that the number of tracks offered by CLEF has increased over the years, from just two in 2000 to nine separate tracks in 2009. As you can note from Figure 4 three tracks have stopped in CLEF 2009 – namely Domain-specific, WebCLEF, and GeoCLEF – and three new tracks have been introduced – namely, LogCLEF, CLEF-IP, and Grid@CLEF. You will find a more detailed description of the CLEF 2009 tracks in the next section.

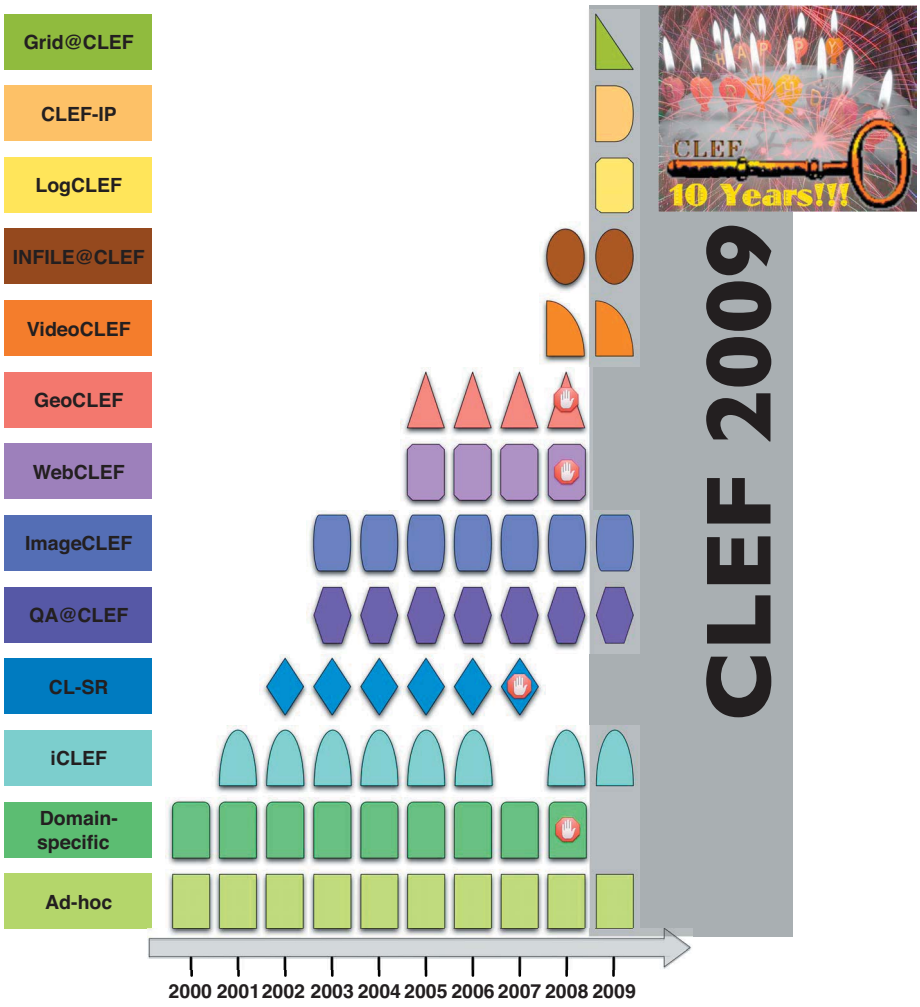

Figure 4: CLEF 2000–2009 tracks.

Each track is run by a coordinating group with specific expertise in the area covered by the track[2].

Most tracks offer several different tasks and these tasks normally vary each year, according to the interests of the track coordinators and participants.The growth in tracks has resulted in a rise in participants; with one exception, the number of participating groups has increased every year. This can be seen in Figure 2 which shows the growth in participation by continent, while Figure 3 shows the participation track by track. Note that many groups participate in more than one track.

Full details of the activities and results of each track, year by year, can be found on the CLEF website in the working notes which are produced at the end of each campaign and which contain detailed reports of the experiments of all participating as well as in the proceedings [9, 10]. In the next section, we comment on the tracks offered in CLEF 2009.

[2]It is impossible to acknowledge all the research organisations that have been involved in the coordination of CLEF from 2000 to 2009. A complete list can be found on the homepage of the CLEF 2000–2009 campaigns at `http://www.clef-campaign.org/`.

2.1 CLEF 2009 Tracks

Multilingual Textual Document Retrieval (Ad Hoc).

The aim of this track is to promote the development of monolingual and cross-language textual document retrieval systems. From 2000–2007, the track exclusively used collections of European newspaper and news agency documents. Last year the focus of the track was considerably widened: we introduced very different document collections, a non-European target language, and an IR task designed to attract participation from groups interested in *Natural Language Processing (NLP)*. The 2009 Ad Hoc track was to a large extent a repetition of last year's track, with the same three tasks: Tel@CLEF, Persian@CLEF, and Robust-WSD. An important objective has been to create good reusable test collections for each of them The track was thus structured in three distinct streams.

The first task offered monolingual and cross-language search on library catalog records and was organized in collaboration with *The European Library (TEL)*[3]. The second task resembled the ad hoc retrieval tasks of previous years but

[3]`http://www.theeuropeanlibrary.org/`

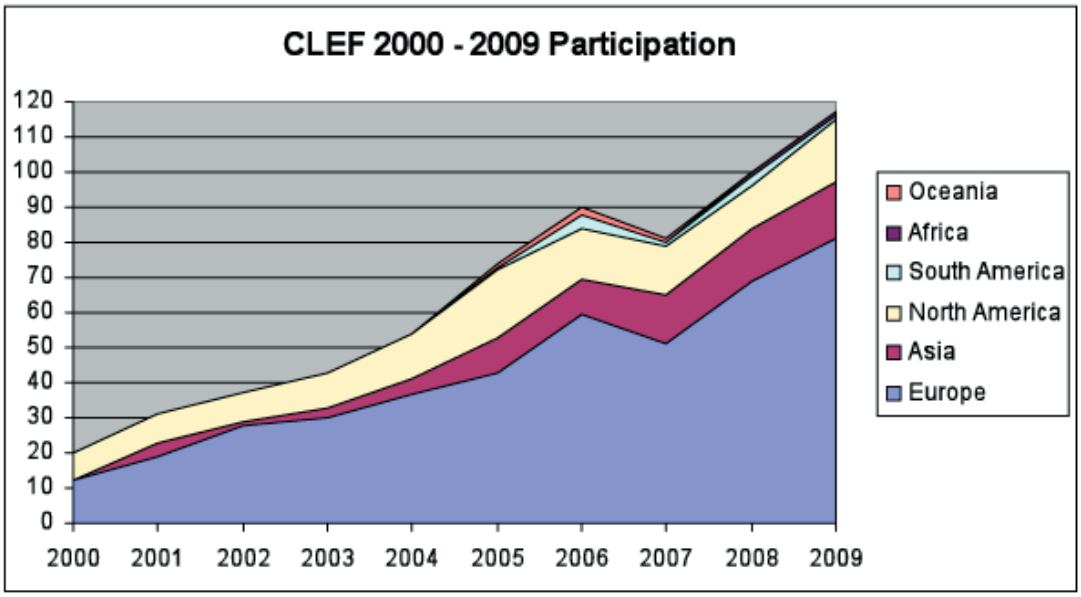

Figure 5: CLEF 2000–2009 participation.

this time the target collection was a Persian newspaper corpora. The third task was the robust activity which used *Word Sense Disambiguated (WSD)* data. The track was coordinated jointly by ISTI-CNR and Padua University, Italy; the University of the Basque Country, Spain; with the collaboration of the Database Research Group, University of Tehran, Iran.

Interactive Cross-Language Retrieval (iCLEF).

In iCLEF, cross-language search capabilities are studied from a user-inclusive perspective. A central research question is how best to assist users when searching information written in unknown languages, rather than how best an algorithm can find information written in languages different from the query language. Since 2006, iCLEF has based its experiments on Flickr, a large-scale, web-based image database where image annotations constitute a naturally multilingual folksonomy.

In an attempt to encourage greater participation in user-orientated experiments, a new task was designed for 2008 and has had a continuation in 2009. The main novelty has been to focus experiments on a shared analysis of a large search log, generated by iCLEF participants from a single search interface provided by the iCLEF organizers. The focus is, therefore, on search log analysis rather than on system design. The idea is to study the behaviour of users in an (almost) naturalistic search scenario, having a much larger data set than in previous iCLEF campaigns. The track was coordinated by UNED, Madrid, Spain; Sheffield University, UK; Swedish Institute of Computer Science, Sweden.

Multilingual Question Answering (QA@CLEF).

This track has offered monolingual and cross-language question answering tasks since 2003. QA@CLEF 2009 proposed three exercises: ResPubliQA, QAST and GikiCLEF:

- **ResPubliQA**: The hypothetical user considered for this exercise is a person close to the law domain interested in making inquiries on European legislation. Given a pool of 500 independent natural language questions, systems must return the passage that answers each question (not the exact answer) from the JRC-Acquis collection of EU parliamentary documentation. Both questions and documents are translated and aligned for a subset of languages. Participating systems could perform the task in Basque, Bulgarian, English, French, German, Italian, Portuguese, Romanian and Spanish.
- **QAST**: The aim of the third QAST exercise was to evaluate QA technology in a real multilingual speech scenario in which written and oral questions (factual and definitional) in different languages are formulated against a set of manually and automatically transcribed audio recordings related to speech events in those languages. The scenario proposed was the European Parliament sessions in English, Spanish and French.
- **GikiCLEF**: Following the previous GikiP pilot at GeoCLEF 2008, the task focused on open list questions over Wikipedia that require geographic reasoning, complex information extraction, and cross-lingual processing, for collections in Bulgarian, Dutch, English, German, Italian, Norwegian (both BokmN̆l and Nynorsk), Portuguese and Romanian or Spanish.

The track was organized by a number of institutions (one for each target language), and jointly coordinated by CELCT, Trento, Italy, and UNED, Madrid, Spain.

Cross-Language Retrieval in Image Collections (ImageCLEF).

This track evaluated retrieval from visual collections; both text and visual retrieval techniques were employed. A number of challenging tasks were offered:

- multilingual ad-hoc retrieval from a photo collection concentrating on diversity in the results;

- a photographic annotation task using a simple ontology;
- retrieval from a large scale, heterogeneous collection of Wikipedia images with user-generated textual metadata, and queries in several languages;
- medical image retrieval (with visual, semantic and mixed topics in several languages);
- medical image annotation;
- detection of semantic categories from robotic images (non-annotated collection, concepts to be detected).

A large number of organisations have been involved in the complex coordination of these tasks. They include: Sheffield University, UK; University of Applied Sciences Western Switzerland; Oregon Health and Science University, USA; University of Geneva, Switzerland; CWI, The Netherlands; IDIAP, Switzerland. The ImageCLEF track has organised a separate one day workshop on visual information retrieval evaluation in collaboration with the Theseus project.

Multilingual Information Filtering (INFILE@CLEF).
INFILE (INformation, FILtering & Evaluation) is a cross-language adaptive filtering evaluation track sponsored by the French National Research Agency. INFILE has extended the last filtering track of TREC 2002 as follows. It uses a corpus of 100,000 Agence France Press comparable newswires for Arabic, English and French; evaluation is performed using an automatic querying of test systems with a simulated user feedback. Each system can use the feedback at any time to increase performance.

The track has been coordinated by the Evaluation and Language resources Distribution Agency (ELDA), France; University of Lille, France; and CEA LIST, France.

Cross-Language Video Retrieval (VideoCLEF).
VideoCLEF 2009 is dedicated to developing and evaluating tasks involving access to video content in a multilingual environment. Participants were provided with a corpus of video data (Dutch-language television, predominantly documentaries) accompanied by speech recognition transcripts. In 2009, there were three tasks: "Subject Classification", which involved automatically tagging videos with subject labels; "Affect", which involved classifying videos according to characteristics beyond their semantic content; "Finding Related Resources Across Languages", which involved linking video to material on the same subject in a different language.

The track was jointly coordinated by Delft University of Technology and Dublin City University, Ireland.

Intellectual Property (CLEF-IP).
This was the first year for the CLEF-IP track. The purpose of the track was twofold: to encourage and facilitate research in the area of patent retrieval by providing a large clean data set for experimentation; to create a large test collection of patents in the three main European languages for the evaluation of cross-lingual information access. The track focused on the task of prior art search. A large test collection for evaluation purposes was created by exploiting patent citations. The collection consists of a corpus of 1,9 million patent documents and 10,000 topics with an average of 6 relevance assessments per topic.

The track was jointly coordinated by the Information Retrieval Facility (IRF), Austria, and Matrixware, Austria.

Log file analysis (LogCLEF).
LogCLEF is an evaluation initiative for the analysis of queries and other logged activities as expression of user behaviour. The goal is the analysis and classification of queries in order to understand search behaviour in multilingual contexts and ultimately to improve search systems. The track used log data from the files of The European Library.

The track was jointly coordinated by Delft University of Hildesheim, Germany, and University of Padua, Italy.

Grid Experiments (Grid@CLEF).
This experimental pilot is planned as a long term activity with the aim of: looking at differences across a wide set of languages; identifying best practices for each language; helping other countries to develop their expertise in the IR field and create IR groups. Participants had to conduct experiments according to the CIRCO (Coordinated Information Retrieval Components Orchestration) protocol, an XML-based framework which allows for a distributed, loosely-coupled, and asynchronous experimental evaluation of Information Retrieval (IR) systems.

The track was coordinated jointly by University of Padua, Italy, and the National Institute of Standards and Technology, USA.

2.2 Main Results

In its first ten year of activities, CLEF has played a leading role in stimulating the investigation and research in a wide range of key areas, such as cross-language question answering, image and geographic information retrieval, interactive retrieval and many more. Moreover, it promoted the study and implementation of appropriate evaluation methodologies for these diverse types of tasks and media.

All the activities led to the creation of well-known and reusable test collections[4] which allow researchers to adopt a comparative evaluation approach in which system performances are compared according to the Cranfield methodology [3]. The CLEF test collections are thus made up of multilingual and multimedia documents, topics and relevance assessments.

Moreover, the different tracks and tasks produced a vast amount of valuable scientific data, resulting from their benchmarking activities, that allow researchers and developers to derive quantitative and qualitative evidence with respect to best practice in multilingual and multimedia information system development.

These data are now managed and made accessible by means of the *Distributed Information Retrieval Evaluation Campaign Tool (DIRECT)* system [2], which keeps the data yearly produced during an evaluation campaign and support all the tasks needed in an evaluation campaign, such as topic creation, experiment submission, pooling and relevance assessments, performance measure computation, and so on. DIRECT now manages: more than 5.6 million documents; more than 1.5 million relevance assessments for more than 600 topics made by over 200 assessors in 15 countries; more than 2,500 experiments, which amount to about 117 million

[4]The 2000-2008 test collections are now publicly available on the *Evaluation and Language resources Distribution Agency (ELDA)* catalog, see `http://www.elda.org/`.

tuples, submitted by over 170 participants in about 30 different countries; over 5.5 million performance measures and descriptive statistics; and, about 20,000 plots and statistical analyses graphs.

Finally, CLEF has been extremely successful in building a wide, strong, and multidisciplinary research community, which covers and spans the different expertises needed to deal with the spread of CLEF tracks and tasks. This research community, which has been constantly growing over the years, has put, almost completely on a volunteer basis, an incredible amount of effort in making CLEF happen and it is at the core of CLEF achievements.

3. CLEF 2010

CLEF 2010[5] represents a renewal of the "classic CLEF" format and an experiment to understand how "next generation" evaluation campaigns might be structured. We had to face the problem of how to innovate CLEF still preserving its traditional core business, namely the benchmarking activities carried out in the various tracks and tasks.

The result of lively and community-wide discussions has been to make CLEF an independent four days event[6] constituted by two main parts: a peer-reviewed conference – the first two days – and a series peer-reviewed laboratories and workshop – the second two days.

The conference part aims at advancing the evaluation of complex multimodal and multilingual information systems in order to support individuals, organizations, and communities who design, develop, employ, and improve such systems. Scientific papers have been solicited in order to explore needs and practices for information access; study new evaluation metrics and methodologies; discuss new directions for future activities in the European multilingual and multimodal IR system evaluation in context; and, analyse achievements in 10 years of CLEF by means of in-depth experiments using existing CLEF collections in any imaginable and interesting way. A large programme committee, representative not only of the multidisciplinary competencies which have traditionally been part of CLEF but also covering new areas, has been established to stimulate papers on the following topics:

- novel methodologies for the design of evaluation tasks, especially user-centric ones;
- analysis of the impact of multilingual, multicultural, multimodal differences in interface and search design;
- assessing multilinguality and multimodality in relevant application communities, e.g. digital libraries, intellectual property, medical, music, video, and social media.
- alternative methods for improving and automating the creation of ground-truth, for example crowd-sourcing or clicklog-based;
- prediction of success and satisfaction rate;
- task-oriented metrics of success and failure;
- evaluation of technology vs testing of scientific theories;
- innovative and easy to communicate techniques for analysing the experimental results, including statistical analyses, data mining, and information visualization;
- alternatives and comparison of item-based, list-based, set-based, and session-based evaluation;
- simulation (of queries, sessions, users) and information retrieval;
- infrastructures for bringing automation and collaboration in the evaluation process;
- component-based evaluation approaches;
- evaluation and analysis using private or anonymized test data;
- living laboratories and evaluating live systems.

In addition, to stimulate the exploitation, re-use, and deep analysis of ten year of CLEF data, we have made them freely available online, upon registration [1]. Figure 6 shows a screenshot of the interface for accessing the whole set of scientific data produced during the history of CLEF. On the left, there is a tree which allows the user to browse thorough the CLEF campaigns from 2000 to 2009 and, for each campaign, it is possible to see what tracks and tasks are available and download all the related data and information.

The laboratories continue and improve the tracks which are traditional in CLEF. Two different forms of labs are offered: benchmarking activities which are very similar to the CLEF tracks evaluation campaigns, and workshop-style labs that explore issues of information access evaluation and related fields. Labs have to fulfill some selection criteria, such as soundness of methodology, feasibility of task; use case, business case/industrial stakeholders; number of potential participants; clear movement along a growth path, scale of experiments; reusability, minimize overlap with other campaigns and labs, interdisciplinary; and so on. A lab selection committee has been established in order to peer-review lab proposals and decide on which to accept for CLEF 2010. The objective of this new procedure is twofold: (i) to try to address a long-standing issue in CLEF, i.e. tracks which are never ending due to their enthusiastic volunteering basis, by ensuring a fair and commonly understood review process; (ii) to try to make the benchmarking activities as adherent as possible to the challenges and scenario envisioned in Section 1.

CLEF 2010 offers five labs and 2 workshops. The selected labs are:

CLEF-IP puts to use a collection of almost 2 million patent documents in *eXtensible Markup Language (XML)* format with content in English, German, and French. The lab offers a *Prior Art Candidate Search* task and a *Classification* task. The first task will ask participants to retrieve documents that are potential prior art to a given document. Topics will be chosen as to stimulate multilingual retrieval. The second task will ask participants to classify documents according to the International Patent Classification scheme. Training data for

[5] `http://www.clef2010.org/`

[6] So far, CLEF has been running as a two days and half workshop in conjunction with the *European Conference on Research and Advanced Technology for Digital Libraries (ECDL)*.

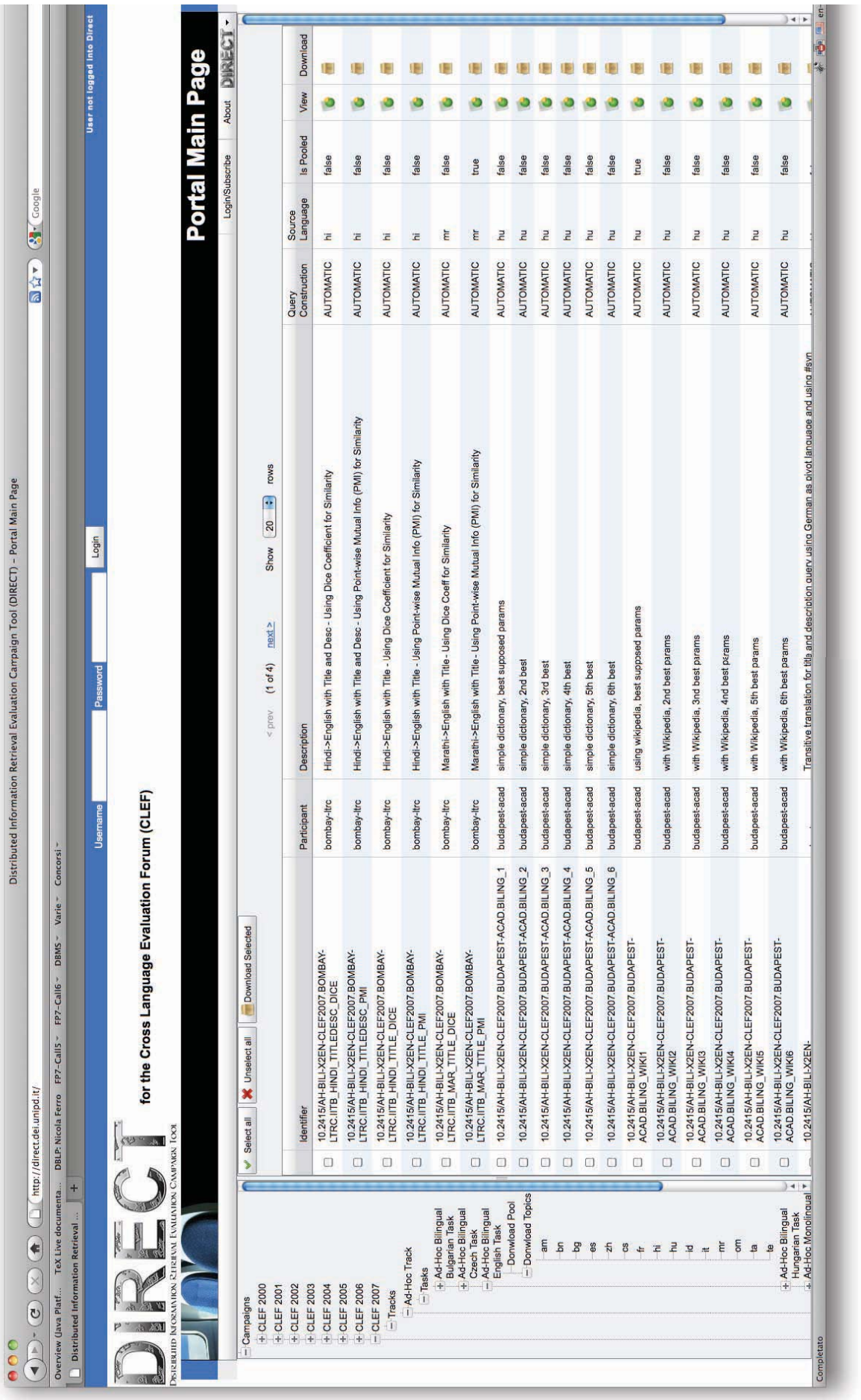

Figure 6: DIRECT interface for accessing the history of ten years of CLEF data.

both tasks will be available prior to the topic sets release. Relevance assessment will be done using patent citations for the first task and current patent classifications for the second task. Coordinator: Information Retrieval Facility (IRF), Austria.

Cross-Language Image Retrieval (ImageCLEF) This labs evaluates retrieval from visual collections; both text and visual retrieval techniques are exploitable. Four challenging tasks are foreseen: 1) retrieval from a Wikipedia collection containing images and structured information in several languages; 2) medical image retrieval with visual, semantic and mixed topics in several languages with a data collection from the scientific literature; 3) detection of semantic categories from robotic images (non-annotated collection, concepts to be detected); 4) a photo annotation task that investigates automated semantic annotation based on visual information with approaches based on Flickr user tags and multimodal approaches. Track coordinators are U. of Applied Sciences Western Switzerland (CH), Oregon Health and Science U. (US), CWI (NL), TELECOM Bretagne (FR), Leiden University (NL), U. of Geneva (CH), Fraunhofer Society (DE), IDIAP (CH).

Uncovering Plagiarism, Authorship, and Wikipedia Vandalism (PAN) [New this year] This lab divides into two tasks:

- *Plagiarism Detection.* Today's plagiarism detection systems are faced with intricate situations, such as obfuscated plagiarism or plagiarism within and across languages. Moreover, the source of a plagiarism case may be hidden in a large collection of documents, or it may not be available at all. Following the success of the 2009 campaign on plagiarism detection, we will provide a revised evaluation corpus consisting of artificial and simulated plagiarism.
- *Wikipedia Vandalism Detection.* Vandalism has always been one of Wikipedia's biggest problems. However, the detection of vandalism is done mostly manually by volunteers, and research on automatic vandalism detection is still in its infancy. Hence, solutions are to be developed which aid Wikipedians in their efforts.

The lab is organized by the Bauhaus-UniversitŁt Weimar, the Universidad PolitŐcnica de Valencia, the University of the Aegean, and the Bar-Ilan University

ResPubliQA Two separate tasks are proposed for the ResPubliQA 2010 evaluation campaign which allow both passages and exact answers (smallest exact demarcation) to be returned as the type of answer in response to the same 200 input questions. Systems can also return NOA if they are not confident of their answer. The focus is on the direct comparison of systems' performances among languages, a goal which is enabled by the adoption of the multilingual parallel paragraph-aligned document collections (JRC-Acquis and Europarl) of EU legislative documents, available in 9 languages, (i.e: Bulgarian, Dutch, English, French, German, Italian, Portuguese, Romanian and Spanish). The Lab is jointly coordinated by UNED, CELCT and the University of Limerick.

WePS [New this year] is a competitive evaluation campaign which consists of two tasks concerning the Web Entity Search Problem: - Task 1 is related to Web People Search, and focuses on person name ambiguity and person attribute extraction on Web pages. Given a set of web search results for a person name, the tasks consists of clustering the pages according to the different people sharing the name and extract certain biographical attributes for each person. - Task 2 is related to Online Reputation Management for organizations, and focuses on the problem of ambiguity for organization names and the relevance of Web data for reputation management purposes. Given a set of Twitter entries containing an (ambiguous) company name, and given the home page of the company, the tasks consists of discriminating entries that do not refer to the company. WePS 3 is coordinated by three universities (UNED, New York University, the University of Illinois at Chicago) and two corporate stakeholders: Intelius Corp. and Llorente & Cuenca.

The selected workshops are:

Cross-lingual Expert Search - Bridging CLIR and Social Media (CriES) [New this year] It addresses the problem of multilingual expert search in social media environments. The main topics are multilingual expert retrieval methods, social media analysis with respect to expert search, selection of data sets and evaluation of expert search results. In addition to the workshop we also organize a pilot challenge:(i) *Workshop*: We expect submissions addressing the main topics including user characterization in multi-lingual social media, community analysis for retrieval scenarios, user-centric recommender algorithms, proposals of new social media datasets and evaluation of cross-lingual expert search. (ii) *Pilot Challenge*: The challenge is based on a dataset from Yahoo!Answers, consisting of multi-lingual questions, answers and user relations. Given a set of multi-lingual questions the task is to retrieve relevant users that will most likely be able to answer the questions. Coordinators are KIT, U. of Koblenz and U. of Bielefeld (DE).

LogCLEF The goal of LogCLEF is the analysis and classification of queries in order to understand search behavior in multilingual contexts and ultimately to improve search systems. A common data set will be distributed to the participants. In coordination with the organizers, participating groups will be devoted to different tasks in exploring and understanding the data. Tasks will include the identification of the language of a query, identification of sessions with more than one language, user clustering, labelling named entities (esp. person names and geographic names) and linking them to these entities (e.g. Wikipedia pages). Both search log and HTTP logs for the 2007/2008 (the same period used in 2009), plus the search log of 2009 will be (most likely) available from The European Library. At the workshop, participants are required to present their algorithms, their results and discuss what the results tell about user behavior. The workshop will be

the basis for a definition of a set of competitive tasks for future studies on log analysis at LogCLEF. Coordinators are: University of Hildesheim and University of Padua.

As shown in Figure 7, the overall vision for CLEF 2010 is to be a dynamic and live entity, which will act as a forum where researchers, developers, and stakeholders in the multilingual and multimedia information systems field will have the opportunity to meet, collaborate, share ideas and knowledge, and conduct their own evaluation activities.

CLEF 2010 is characterised by:

- **basement**: the conference aimed at advancing the evaluation of multilingual and multimedia information systems, the evaluation methodologies and metrics developed to embody realistic use cases and evaluation task; and, the techniques aimed at bringing more automation in the evaluation process constitute the basement of CLEF 2010 and will provide the foundations for promoting and supporting the expected scientific and technological advancement.
- **pillars**: the regular and thorough evaluation activities carried out in the labs, the realistic use cases and evaluation tasks designed for compelling user and industrial needs represent the pillars of CLEF 2010. They stimulate the research and development in the multilingual and multimedia information systems field and they contribute to the creation and driving of a multidisciplinary researchers and developers community which brings together the competencies needed to develop such complex systems.
- **roof**: the "basement" and the "pillars" of CLEF 2010 will give the necessary support for designing and developing the next generation multilingual and multimedia information systems needed to address the emerging user needs and to cope with the interaction among content, users, languages and task discussed in Section 1.

4. FUTURE PERSPECTIVES

Our plans, after CLEF 2010, are to continue to advance the experimental evaluation of complex multimedia and multilingual information systems in order to support individuals, commercial entities, and communities who design, develop, employ, and improve such complex systems.

To this end, we will provide a virtual and open laboratory for conducting participative research and experimentation in which it will be possible to carry out, advance and bring automation into the evaluation and benchmarking of complex multimedia and multilingual information systems, by facilitating management and offering access, curation, preservation, re-use, analysis, visualisation, and mining of the collected experimental data.

In order to pursue this goal, we will leverage on the key activities shown in Figure 8:

- **foster the adoption of regular and thorough experimental evaluation activities**: we will rely on and expand the large data sets that have been developed in CLEF over the years; tackle realistic tasks and use cases; advance the evaluation methodologies to better support the realistic and user-centered tasks and use cases; involve large developer and researcher communities; provide a proper evaluation infrastructure to support the evaluation activities; produce a growing knowledge-base where experimental collections, experimental results and evidence, evaluation measures and analyses will accumulate and be available for further study and analysis.
- **bring automation into the experimental evaluation process**: we will propose methods and provide tools for the creation of larger experimental collections; increase the number and size of the experiments conducted; and develop distributed, asynchronous, and loosely-coupled evaluation protocols.
- **promote collaboration and re-use over the acquired knowledge-base**: we will curate, preserve, and enrich the collected experimental data; provide the means for an easy comparison with and a meaningful interpretation and visualisation of the experimental results; and facilitate the discussion and collaboration among all the interested stakeholders.
- **stimulate knowledge transfer and uptake**: we will disseminate know-how, tools, and best practices about multilingual and multimedia information systems; facilitate uptake and participation by commercial entities and industries; and give rise to multidisciplinary competencies and expertises.

The first key activity will be the catalyst for promoting innovation and bringing together multidisciplinary expertises. Indeed, the evaluation activities will be based on well-defined and compelling use cases which will grasp both the different facets of the evolution of multilingual and multimedia information systems and the interaction between content, user, languages, and tasks discussed in Section 1. This will give raise to the need of mixing competencies coming from different areas of expertise and the CLEF will play a fundamental role in providing the forum and the instruments for making this happen.

The second key activity concerns technical aspects of the evaluation practices that need to be improved in order to effectively support the vision discussed in Section 1. In this area, CLEF will introduce for the first time methods and tools for moving the experimental evaluation from an handicraft process to a mostly automatic one.

The third key activity will explore new ways of exploiting the knowledge created by evaluation activities and of actively involving all the stakeholders with the knowledge-base which will be progressively accumulated. In particular, we will pioneer the application of visual analytics [6] techniques to experimental evaluation, as well as the exploitation of active communications means, such as digital annotations and social tagging.

Finally, the fourth key activity is concerned with the spreading of the excellence and the long term integration of the acquired competencies in order to effectively share the obtained achievements at an European level.

5. ACKNOWLEDGMENTS

The author would like to warmly thank Maristella Agosti, Donna Harman, and Carol Peters for their continuous support and advice. The author also thanks Mihai Lupu for the image about multilingual patents.

Figure 7: CLEF 2010 vision.

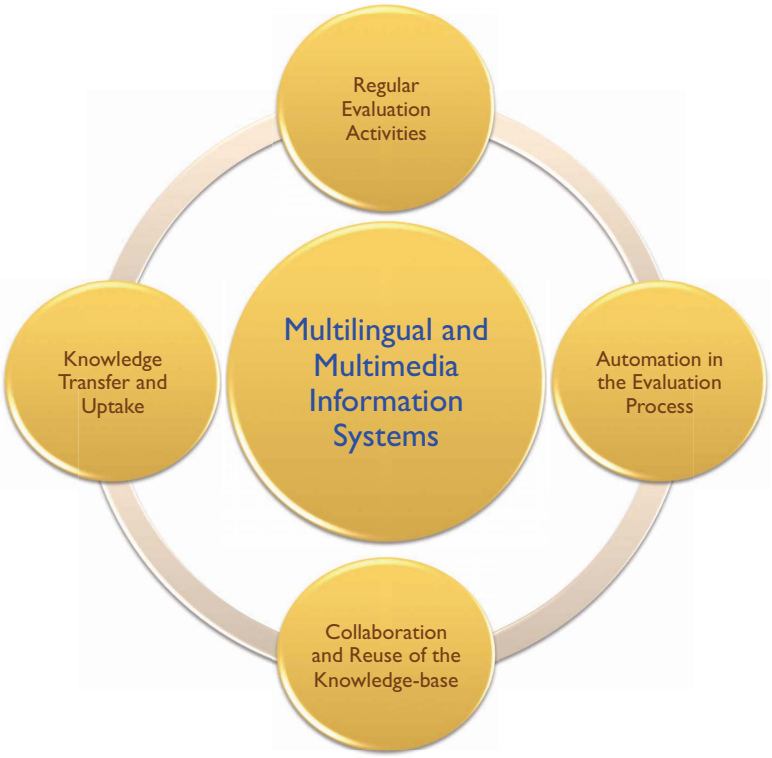

Figure 8: Beyond CLEF 2010.

6. REFERENCES

[1] M. Agosti, G. M. Di Nunzio, M. Dussin, and N. Ferro. 10 Years of CLEF Data in DIRECT: Where We Are and Where We Can Go. In T. Sakay, M. Sanderson, and W. Webber, editors, *Proc. 3rd International Workshop on Evaluating Information Access (EVIA 20010)*. National Institute of Informatics, Tokyo, Japan, 2010.

[2] M. Agosti and N. Ferro. Towards an Evaluation Infrastructure for DL Performance Evaluation. In G. Tsakonas and C. Papatheodorou, editors, *Evaluation of Digital Libraries: An insight into useful applications and methods*, pages 93–120. Chandos Publishing, Oxford, UK, 2009.

[3] C. W. Cleverdon. The Cranfield Tests on Index Languages Devices. In K. Spärck Jones and P. Willett, editors, *Readings in Information Retrieval*, pages 47–60. Morgan Kaufmann Publisher, Inc., San Francisco, CA, USA, 1997.

[4] M. Dussin and N. Ferro. Managing the Knowledge Creation Process of Large-Scale Evaluation Campaigns. In M. Agosti, J. Borbinha, S. Kapidakis, C. Papatheodorou, and G. Tsakonas, editors, *Proc. 13th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2009)*, pages 63–74. Lecture Notes in Computer Science (LNCS) 5714, Springer, Heidelberg, Germany, 2009.

[5] N. Ferro and C. Peters. From CLEF to TrebleCLEF: the Evolution of the Cross-Language Evaluation Forum. In N. Kando and M. Sugimoto, editors, *Proc. 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access*, pages 577–593. National Institute of Informatics, Tokyo, Japan, 2008.

[6] D. A. Keim, F. Mansmann, J. Schneidewind, and H. Ziegler. Challenges in Visual Data Analysis. In E. Banissi, editor, *Proc. of the 10th International Conference on Information Visualization (IV 2006)*, pages 9–16. IEEE Computer Society, Los Alamitos, CA, USA, 2006.

[7] H. Müller, J. Kalpathy-Cramer, C. E. Kahn, W. Hatt, S. Bedrick, and W. Hersh. Overview of the ImageCLEFmed 2008 Medical image Retrieval Task. In C. Peters, T. Deselaers, N. Ferro, J. Gonzalo, G. J. F. Jones, M. Kurimo, T. Mandl, and A. Peñas, editors, *Evaluating Systems for Multilingual and Multimodal Information Access: Ninth Workshop of the Cross–Language Evaluation Forum (CLEF 2008). Revised Selected Papers*, pages 500–510. Lecture Notes in Computer Science (LNCS) 5706, Springer, Heidelberg, Germany, 2009.

[8] C. Peters. Introduction. In C. Peters, editor, *Cross-Language Information Retrieval and Evaluation: Workshop of Cross-Language Evaluation Forum (CLEF 2000)*, pages 1–6. Lecture Notes in Computer Science (LNCS) 2069, Springer, Heidelberg, Germany, 2001.

[9] C. Peters, G. M. Di Nunzio, M. Kurimo, T. Mandl, D. Mostefa, A. Peñas, and G. Roda, editors. *Multilingual Information Access Evaluation Vol. I Text Retrieval Experiments – Tenth Workshop of the Cross–Language Evaluation Forum (CLEF 2009). Revised Selected Papers*. Lecture Notes in Computer Science (LNCS), Springer, Heidelberg, Germany, 2010.

[10] C. Peters, T. Tsikrika, H. Müller, J. Kalpathy-Cramer, G. J. F. Jones, J. Gonzalo, and B. Caputo, editors. *Multilingual Information Access Evaluation Vol. II Multimedia Experiments – Tenth Workshop of the Cross–Language Evaluation Forum (CLEF 2009). Revised Selected Papers*. Lecture Notes in Computer Science (LNCS), Springer, Heidelberg, Germany, 2010.

[11] J. Tait, M. Lupu, H. Berger, G. Roda, M. Dittenbach, A. Pesenhofer, E. Graf, and C. J. Van Rijsbergen. Patent Search: An important new test bed for IR. In R. Aly, C. Hauff, I. den Hamer, D. Hiemstra, T. Huibers, and F. de Jong, editors, *Proc. 9th Dutch-Belgian Information Retrieval Workshop (DIR 2009)*, pages 56–63. Enschede, University of Twente, Centre for Telematics and Information Technology, 2009.