

Interactive Probabilistic Search for GikiCLEF

Ray R. Larson
School of Information
University of California, Berkeley, USA
ray@sims.berkeley.edu

Abstract

In this paper we will briefly describe the approaches taken by the Berkeley Cheshire Group for the GikiCLEF task of the QA track. Because the task was intended to model some aspects of user search, and because of the complexity of the topics and their geographic elements, we decided to conduct interactive searching of the topics and selection of results. Because of the vagueness of the task specification early-on, some disagreements about what constituted a correct answer, and time constraints we were able to complete only 22 of the 50 topics. However, in spite of this limited submission the interactive approach was very effective and resulted in our submission being ranked third overall in the results.

Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval

General Terms

Algorithms, Performance, Measurement

Keywords

Cheshire II, Logistic Regression

1 Introduction

The GikiCLEF task description, according to the web site is:

“For GikiCLEF, systems will need to answer or address geographically challenging topics, on the Wikipedia collections, returning Wikipedia document titles as list of answers.”

“The user model for which GikiCLEF systems intend to cater for is anyone who is interested in knowing something that might be already included in Wikipedia, but has not enough time or imagination to browse it manually.”

“So, in practice, a system participating in GikiCLEF receives a set of topics – representing valid and realistic user needs preferably from non-English users – in all GikiCLEF languages and will have to produce a list of answers, in all languages it can find answers.”

“The motivation for this kind of system behaviour is that in a real environment, a post-processing module For different kinds of human users, and depending on the languages those users could read, different possible output formats would filter the information per language, or would rank it in order of preference. We are assuming that people prefers to read answers in their native languages, but that many people are happy with answers (answers are titles of Wikipedia entries)

in other languages they also know or even just slightly understand.” (from the GikiCLEF Web site: <http://www.linguateca.pt/GikiCLEF/>)

It was clear from the description that some form of interactive search was intended, but the task description is not at all clear on what constitutes an “answer” to a particular question. Because we did not know enough about the task and to attempt a fully automated approach, we decided to construct an interactive IR system that was able to search across all of the Wikipedia test collections in each language (Bulgarian, Dutch, English, German, Italian, Norwegian (bokmaal), Norwegian (nynorsk), Portuguese, Romanian and Spanish)

What was not clear was that the intended answers to the questions could *not be in the text of the articles* but had to be exactly the *title of the article*, and that title had to be of a specific type (often a place) that was supposed to be inferred from the form of the question. This constraint effectively eliminated the possibility for fully automatic methods (at least given the techniques we had readily available), so we decided to use this first participation in GikiCLEF as an exploratory study of what kinds of search might prove useful in this task

In this paper we will describe the retrieval algorithms employed in our interactive system, some description of the system itself, and some comments on the evaluation and various issues that arose.

2 The Retrieval Algorithms

Note that much of this section is virtually identical to one that appears in our papers from previous CLEF participation[9, 8] The retrieval algorithms used for our GikiCLEF interactive system include ranked retrieval using our Logistic Regression algorithm combined with Boolean constraints, as well as simple Boolean queries for link following, etc.

The basic form and variables of the *Logistic Regression* (LR) algorithm used for all of our submissions was originally developed by Cooper, et al. [6]. As originally formulated, the LR model of probabilistic IR attempts to estimate the probability of relevance for each document based on a set of statistics about a document collection and a set of queries in combination with a set of weighting coefficients for those statistics. The statistics to be used and the values of the coefficients are obtained from regression analysis of a sample of a collection (or similar test collection) for some set of queries where relevance and non-relevance has been determined. More formally, given a particular query and a particular document in a collection $P(R | Q, D)$ is calculated and the documents or components are presented to the user ranked in order of decreasing values of that probability. To avoid invalid probability values, the usual calculation of $P(R | Q, D)$ uses the “log odds” of relevance given a set of S statistics, s_i , derived from the query and database, such that:

$$\log O(R | Q, D) = b_0 + \sum_{i=1}^S b_i s_i \quad (1)$$

where b_0 is the intercept term and the b_i are the coefficients obtained from the regression analysis of the sample collection and relevance judgements. The final ranking is determined by the conversion of the log odds form to probabilities:

$$P(R | Q, D) = \frac{e^{\log O(R|Q,D)}}{1 + e^{\log O(R|Q,D)}} \quad (2)$$

2.1 TREC2 Logistic Regression Algorithm

For GikiCLEF we used a version the Logistic Regression (LR) algorithm that has been used very successfully in Cross-Language IR by Berkeley researchers for a number of years[3]. The formal definition of the TREC2 Logistic Regression algorithm used is:

$$\log O(R|C, Q) = \log \frac{p(R|C, Q)}{1 - p(R|C, Q)} = \log \frac{p(R|C, Q)}{p(\bar{R}|C, Q)}$$

$$\begin{aligned}
&= c_0 + c_1 * \frac{1}{\sqrt{|Q_c|} + 1} \sum_{i=1}^{|Q_c|} \frac{qt f_i}{ql + 35} \\
&+ c_2 * \frac{1}{\sqrt{|Q_c|} + 1} \sum_{i=1}^{|Q_c|} \log \frac{t f_i}{cl + 80} \\
&- c_3 * \frac{1}{\sqrt{|Q_c|} + 1} \sum_{i=1}^{|Q_c|} \log \frac{ct f_i}{N_t} \\
&+ c_4 * |Q_c|
\end{aligned} \tag{3}$$

where C denotes a document component (i.e., an indexed part of a document which may be the entire document) and Q a query, R is a relevance variable,

$p(R|C, Q)$ is the probability that document component C is relevant to query Q ,

$p(\bar{R}|C, Q)$ the probability that document component C is *not relevant* to query Q , which is $1.0 - p(R|C, Q)$

$|Q_c|$ is the number of matching terms between a document component and a query,

$qt f_i$ is the within-query frequency of the i th matching term,

$t f_i$ is the within-document frequency of the i th matching term,

$ct f_i$ is the occurrence frequency in a collection of the i th matching term,

ql is query length (i.e., number of terms in a query like $|Q|$ for non-feedback situations),

cl is component length (i.e., number of terms in a component), and

N_t is collection length (i.e., number of terms in a test collection).

c_k are the k coefficients obtained though the regression analysis.

If stopwords are removed from indexing, then ql , cl , and N_t are the query length, document length, and collection length, respectively. If the query terms are re-weighted (in feedback, for example), then $qt f_i$ is no longer the original term frequency, but the new weight, and ql is the sum of the new weight values for the query terms. Note that, unlike the document and collection lengths, query length is the “optimized” relative frequency without first taking the log over the matching terms.

The coefficients were determined by fitting the logistic regression model specified in $\log O(R|C, Q)$ to TREC training data using a statistical software package. The coefficients, c_k , used for our official runs are the same as those described by Chen[1]. These were: $c_0 = -3.51$, $c_1 = 37.4$, $c_2 = 0.330$, $c_3 = 0.1937$ and $c_4 = 0.0929$.

2.2 TREC3 Logistic Regression Algorithm

A variant of the TREC2 algorithm above was also used (or at least was available for use in the interactive GikiCLEF system developed. It does not use the rather arbitrary document length and query length constants used in the TREC2 algorithm, and instead just relies on the regression analysis to provide appropriate weights for the length elements. This equation is essentially the same as that used in by Cooper, et al. [5] in TREC3.

The full equation describing the TREC3 LR algorithm as used in these experiments is:

$$\begin{aligned}
\log O(R | Q, C) &= \\
&b_0 + \left(b_1 \cdot \left(\frac{1}{|Q_c|} \sum_{j=1}^{|Q_c|} \log qt f_j \right) \right)
\end{aligned}$$

$$\begin{aligned}
& + \left(b_2 \cdot \sqrt{|Q|} \right) \\
& + \left(b_3 \cdot \left(\frac{1}{|Q_c|} \sum_{j=1}^{|Q_c|} \log tf_j \right) \right) \\
& + \left(b_4 \cdot \sqrt{cl} \right) \\
& + \left(b_5 \cdot \left(\frac{1}{|Q_c|} \sum_{j=1}^{|Q_c|} \log \frac{N - n_{t_j}}{n_{t_j}} \right) \right) \\
& + \left(b_6 \cdot \log |Q_d| \right)
\end{aligned} \tag{4}$$

Where:

Q is a query containing terms T ,

$|Q|$ is the total number of terms in Q ,

$|Q_c|$ is the number of terms in Q that also occur in the document component,

tf_j is the frequency of the j th term in a specific document component,

qtf_j is the frequency of the j th term in Q ,

n_{t_j} is the number of components (of a given type) containing the j th term,

cl is the document component length measured in bytes.

N is the number of components of a given type in the collection.

b_i are the coefficients obtained through the regression analysis.

The coefficients used were $b_0 = -3.70$, $b_1 = 1.269$, $b_2 = -0.310$, $b_3 = 0.679$, $b_4 = -0.0674$, $b_5 = 0.223$ and $b_6 = 2.01$ for these experiments. Further details on the TREC2 version of the Logistic Regression algorithm may be found in Cooper et al. [4].

2.3 Blind Relevance Feedback

In addition to the direct retrieval of documents using the TREC2 logistic regression algorithm described above, we have implemented a form of “blind relevance feedback” as a supplement to the basic algorithm. The algorithm used for blind feedback was originally developed and described by Chen [2]. Blind relevance feedback has become established in the information retrieval community due to its consistent improvement of initial search results as seen in TREC, CLEF and other retrieval evaluations [7]. The blind feedback algorithm is based on the probabilistic term relevance weighting formula developed by Robertson and Sparck Jones [10].

Blind relevance feedback is typically performed in two stages. First, an initial search using the original topic statement is performed, after which a number of terms are selected from some number of the top-ranked documents (which are presumed to be relevant). The selected terms are then weighted and then merged with the initial query to formulate a new query. Finally the reweighted and expanded query is submitted against the same collection to produce a final ranked list of documents. Obviously there are important choices to be made regarding the number of top-ranked documents to consider, and the number of terms to extract from those documents. For ImageCLEF this year, having no prior data to guide us, we chose to use the top 10 terms from 10 top-ranked documents. The terms were chosen by extracting the document vectors for each of the 10 and computing the Robertson and Sparck Jones term relevance weight for each document. This weight is based on a contingency table where the counts of 4 different conditions for combinations of (assumed) relevance and whether or not the term is, or is not in a document. Table 1 shows this contingency table.

Table 1: Contingency table for term relevance weighting

	Relevant	Not Relevant	
In doc	R_t	$N_t - R_t$	N_t
Not in doc	$R - R_t$	$N - N_t - R + R_t$	$N - N_t$
	R	$N - R$	N

The relevance weight is calculated using the assumption that the first 10 documents are relevant and all others are not. For each term in these documents the following weight is calculated:

$$w_t = \log \frac{\frac{R_t}{R - R_t}}{\frac{N_t - R_t}{N - N_t - R + R_t}} \quad (5)$$

The 10 terms (including those that appeared in the original query) with the highest w_t are selected and added to the original query terms. For the terms not in the original query, the new “term frequency” ($qt f_i$ in main LR equation above) is set to 0.5. Terms that were in the original query, but are not in the top 10 terms are left with their original $qt f_i$. For terms in the top 10 and in the original query the new $qt f_i$ is set to 1.5 times the original $qt f_i$ for the query. The new query is then processed using the same LR algorithm as shown in Equation 4 and the ranked results returned as the response for that topic.

2.4 Boolean Search Operations

To enable efficient browsing in an interactive system it is also useful to implement Boolean constraints and search options. These are built into the Cheshire II system and use the same indexes as the ranked search operations. For this implementation we typically used a Boolean AND search of all of the query words combined by Boolean AND with the results of a ranked search of those words. This operation retains the ranking values generated by the ranked search while limiting the results to only those that contain all of the words in the query.

In addition, to implement the internal links of the Wikipedia test collections for the interactive system, each link in a retrieved page was converted to a Boolean title search for the linked page name. Thus, instead of following links directly each link became a search on a title. Direct use of the links was impossible since the collection pages were not preserved with in the same file structure as the original Wikipedia and names in the links and the actual page file names differed due to additions during the collection creation process.

Also a Boolean AND NOT search was used to help filter results in some cases with ambiguous terms.

3 Approaches for GikiCLEF

In this section we describe the specific approaches taken for our submitted runs for the GikiCLEF task. First we describe the indexing and term extraction methods used, and then the search features we used for the submitted runs.

3.1 Indexing and Term Extraction

The Cheshire II system uses the XML (in this case the XHTML) structure of the documents to extract selected portions for indexing and retrieval. Any combination of tags can be used to define the index contents.

Table 2 lists the indexes created by the Cheshire II system for each language collection of the GikiCLEF Wikipedia database and the document elements from which the contents of those indexes were extracted. For each of the languages: Bulgarian, Dutch, English, German, Italian,

Table 2: Cheshire II Indexes for GikiCLEF 2009

Name	Description	Content Tags
title	Item Title	title tag
meta	Content Metadata	content attribute of meta tag
topic	Most of Record	title, body and meta@content tags
anchors	Anchor text	anchor (a) tags

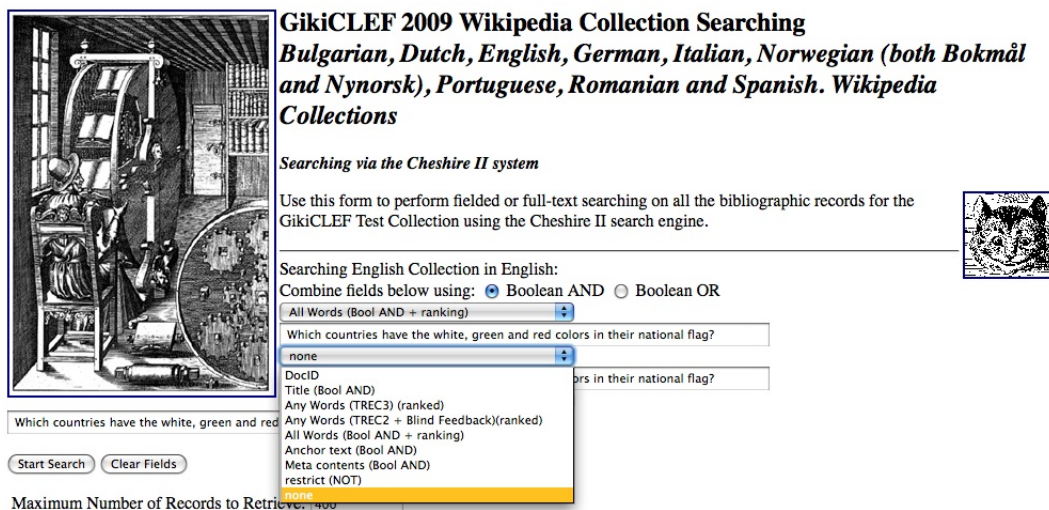


Figure 1: Search Form in the Interactive Search System

Norwegian (bokmaal), Norwegian (nynorsk), Portuguese, Romanian and Spanish, we tried to use, where possible, language-specific stemmers and stoplists whenever possible. Our implementation of the Snowball stemmer is some years old and lacked stemmers for Bulgarian, Norwegian(bokmaal) and Romanian. For these we substituted a stemmer with somewhat similar language roots. I.e., a Russian stemmer for Bulgarian, Norwegian(nynorsk) for Norwegian(Bokmaal) and Italian for Romanian.

For all indexing we used language-specific stoplists to exclude function words and very common words from the indexing and searching. The German language runs *did not* use decomposing in the indexing and querying processes to generate simple word forms from compounds. The Snowball stemmer was used by Cheshire for language-specific stemming.

3.2 Search Processing

Interactive searching of the GikiCLEF Wikipedia collections used the Cheshire II system via a set of web pages and TCL scripts that allowed the searcher to select a particular topic id and language and have it loaded into a search form for manual modification and selection of search indexes and approaches. Figure 1 shows this form for topic GC-2009-02. Typically the user would edit the query to remove extraneous terms, and submit the query, leading to a ranked result list page (Figure 2). From the ranked result list page the user can click on the article title to see the full page (Figure 3) or click on any of the language codes on the line to submit that title as title query in the Wikipedia collection for that language (the user is also given a chance to edit the search before it is submitted to allow language-specific adaptations). From a page like that shown in Figure 3, any of the links can be clicked on generating a Boolean title search for that page. For example clicking on the country name link “Chechnya” in the first line leads to a list of pages



GikiCLEF Test Collection Cheshire II Search Results

Search based on Topic #GC-2009-02 : Which countries have the white, green and red colors in their national flag? (Language = EN)

Your search, encoded as `search(topic {white, green red colors national flag} AND topic @ {white, green red colors national flag})`, is being submitted to the GikiCLEF Test Collection server, where 867 records were found. 400 records will be displayed.

Record #1: Title: [Landesfarben](#) Search in: [BG](#) [DE](#) [EN](#) [ES](#) [IT](#) [NL](#) [NN](#) [NO](#) [PT](#) [RO](#)

Record #2: Title: [List of South African flags](#) Search in: [BG](#) [DE](#) [EN](#) [ES](#) [IT](#) [NL](#) [NN](#) [NO](#) [PT](#) [RO](#)

Record #3: Title: [National Cycling Championships](#) Search in: [BG](#) [DE](#) [EN](#) [ES](#) [IT](#) [NL](#) [NN](#) [NO](#) [PT](#) [RO](#)

Record #4: Title: [Hmar Students Association](#) Search in: [BG](#) [DE](#) [EN](#) [ES](#) [IT](#) [NL](#) [NN](#) [NO](#) [PT](#) [RO](#)

Record #5: Title: [Image:-Burma1300sAvaThu Ye Gyee.jpg](#) Search in: [BG](#) [DE](#) [EN](#) [ES](#) [IT](#) [NL](#) [NN](#) [NO](#) [PT](#) [RO](#)

Record #6: Title: [Flag of Chechnya](#) Search in: [BG](#) [DE](#) [EN](#) [ES](#) [IT](#) [NL](#) [NN](#) [NO](#) [PT](#) [RO](#)

Record #7: Title: [Ghevont Alishan](#) Search in: [BG](#) [DE](#) [EN](#) [ES](#) [IT](#) [NL](#) [NN](#) [NO](#) [PT](#) [RO](#)

Record #8: Title: [George Rogers Clark Flag](#) Search in: [BG](#) [DE](#) [EN](#) [ES](#) [IT](#) [NL](#) [NN](#) [NO](#) [PT](#) [RO](#)

Record #9: Title: [Flag of the Chechen Nation](#) Search in: [BG](#) [DE](#) [EN](#) [ES](#) [IT](#) [NL](#) [NN](#) [NO](#) [PT](#) [RO](#)

Figure 2: Ranked List of Results

containing the word “Chechnya” in their titles, one of which is the specific country page shown in Figure 4.

Each display of a full page includes a “Log as Relevant” button to save the page information in a log file. This log file is the basis for the submitted results for the GikiCLEF task.

4 Results for Submitted Runs

Needless to say, doing the GikiCLEF task interactively involved a lot of time spent reading pages and deciding whether or not the page was relevant. As it turned out in the evaluation many of the pages that I believed to be relevant (such as the page shown in Figure 3 were judged not to be relevant.) Although in this particular case it is very difficult to understand why, for the topic “Which countries have the white, green and red colors in their national flag?” the article entitled “Flag of Chechnya” is considered NOT relevant while the article “Chechnya” IS (even though the colors of the flag are never mentioned and no images were included in the collections). The official position is that the question was about the country, and therefore country names alone are acceptable (the fact that the country name is ALSO included in the non-relevant item does not seem to matter).

In any case, because each question took literally hours of work using the interactive system, and my time was constrained by other considerations (I was on holiday with my family during the time when the topics were available for access and results had to be submitted and could not devote entire days to the task), I completed and submitted only 22 out of the 50 topics, with results from all of the target languages of the collections.

Figure 5 from the GikiCLEF participants web site shows that the interactive approach was fairly effective in spite of not completing all of the topics (our scores are labeled rayrlarson).



GikiCLEF Test Collection Cheshire II Search Results

Your search, encoded as *search docid 5331663*, is being submitted to the *GikiCLEF Test Collection* server, where **1** record was found.

Title Match #1

[Log as Relevant](#)

[BG](#) [DE](#) [EN](#) [ES](#) [IT](#) [NL](#) [NN](#) [NO](#) [PT](#) [RO](#)

Flag of Chechnya

[Chechen Republic flag since 2004](#)

Chechen Republic flag since 2004

The **flag of Chechnya** is a **rectangle** with sides in the ratio 2:3, the same ratio as the flag of the **Russian Federation**. The flag is composed of three horizontal bars of, from top to bottom: **green**, representing **Islam**; white; and **red**; superimposed on them is a narrow vertical **white** band at the hoist side, containing the national ornament, a design of four **golden** scroll shapes.

This flag, introduced in **2004**, is primarily used by the **government** of Chechnya while the independentist flags are commonly used by opposition forces and Chechen people throughout the world.

Historic flags

[Chechen-Ingush ASSR flag in 1957-1978](#)

Chechen-Ingush ASSR flag in 1957-1978

Figure 3: Search Result with Language Search Links

5 Conclusions

In looking at the overall results for the various GikiCLEF tasks, it would appear that the interactive approach using logistic regression ranking Boolean constraints can provide fairly good results. Since GikiCLEF is a new task for us, we took a fairly conservative approach using methods that have worked well in the past, and used our interaction with the collection to try to discover how this kind of searching might be implemented automatically. There are no simple answers for this task with its complex questions and constraints, but through our interactive work we think we have some possible strategies for future evaluation.

References

- [1] Aitao Chen. Multilingual information retrieval using english and chinese queries. In Carol Peters, Martin Braschler, Julio Gonzalo, and Michael Kluck, editors, *Evaluation of Cross-Language Information Retrieval Systems: Second Workshop of the Cross-Language Evaluation Forum, CLEF-2001, Darmstadt, Germany, September 2001*, pages 44–58. Springer Computer Science Series LNCS 2406, 2002.
- [2] Aitao Chen. *Cross-Language Retrieval Experiments at CLEF 2002*, pages 28–48. Springer (LNCS #2785), 2003.
- [3] Aitao Chen and Fredric C. Gey. Multilingual information retrieval using machine translation, relevance feedback and decompounding. *Information Retrieval*, 7:149–182, 2004.
- [4] W. S. Cooper, A. Chen, and F. C. Gey. Full Text Retrieval based on Probabilistic Equations with Coefficients fitted by Logistic Regression. In *Text REtrieval Conference (TREC-2)*, pages 57–66, 1994.
- [5] William S. Cooper, Fredric C. Gey, and Aitao Chen. Full text retrieval based on a probabilistic equation with coefficients fitted by logistic regression. In Donna K. Harman, editor, *The*

[REDIRECT TO: Chechnya](#)

Title Match #26

[Log as Relevant](#)


[BG](#) [DE](#) [EN](#) [ES](#) [IT](#) [NL](#) [NN](#) [NO](#) [PT](#) [RO](#)

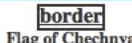
Chechnya

Chechen Republic (English)
Чеченская Республика (Russian)
Нохчийн Республика (Chechen)

|
Location of the Chechen Republic in Russia

[Coat of Arms](#)[Flag](#)

Coat of arms


border


Anthem: [Anthem of the Chechen Republic](#)

Capital	Grozny
Established	January 11 , 1991
Political status	Republic
Federal district	Southern
Economic region	North Caucasus
Code	20

Area

Area	15,300 km²
- Rank within Russia	75th

Population

(as of the 2002 Census)

Figure 4: Multilingual Results with Various Search Links

Second Text Retrieval Conference (TREC-2) (NIST Special Publication 500-215), pages 57–66, Gaithersburg, MD, 1994. National Institute of Standards and Technology.

- [6] William S. Cooper, Fredric C. Gey, and Daniel P. Dabney. Probabilistic retrieval based on staged logistic regression. In *15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Copenhagen, Denmark, June 21-24*, pages 198–210, New York, 1992. ACM.
- [7] Ray R. Larson. Probabilistic retrieval, component fusion and blind feedback for XML retrieval. In *INEX 2005*, pages 225–239. Springer (Lecture Notes in Computer Science, LNCS 3977), 2006.
- [8] Ray R. Larson. Cheshire at geoclef 2007: Retesting text retrieval baselines. In *8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers*, LNCS 5152, pages 811–814, Budapest, Hungary, September 2008.
- [9] Ray R. Larson. Experiments in classification clustering and thesaurus expansion for domain specific cross-language retrieval. In *8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers*, LNCS 5152, pages 188–195, Budapest, Hungary, September 2008.

Final score

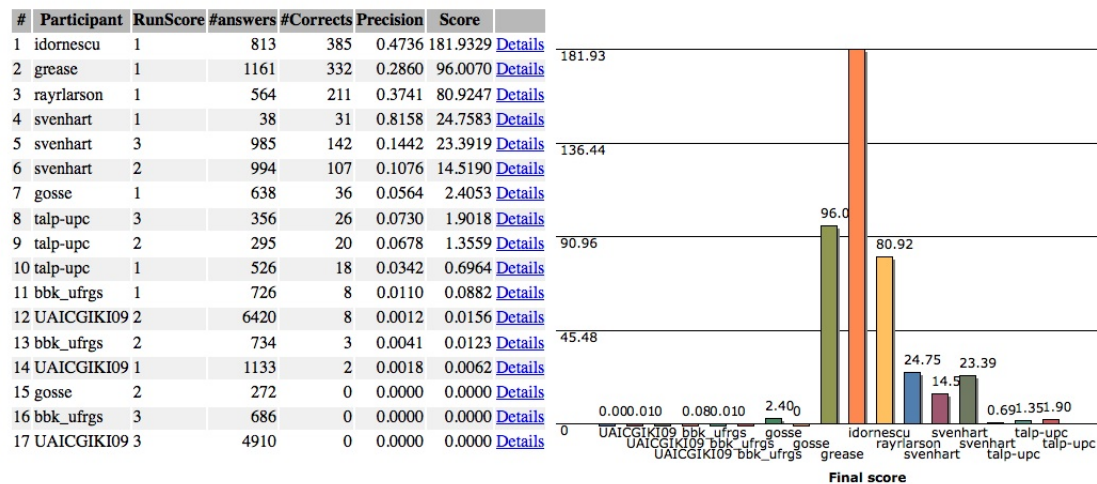


Figure 5: Final Evaluation Scores for GikiCLEF

- [10] S. E. Robertson and K. Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, pages 129–146, May–June 1976.