PatOlympics - An Infrastructure for Interactive Evaluation of Patent Retrieval Tools

Mihai Lupu*
Information Retrieval Facility, and
Vienna University of Technology
lupu@ifs.tuwien.ac.at

ABSTRACT

We present PatOlympics - the interactive evaluation campaign organized by the Information Retrieval Facility in the context of its yearly symposium. In particular, we focus in this paper on the infrastructure behind the event. This infrastructure, consisting of a relational database back-end, a Java processing core and JavaScript interface, makes it possible for real users and IR researchers to interact in an exciting, competitive environment, while maintaining, to the extent possible, the evaluation procedures of standard IR campaigns.

Categories and Subject Descriptors

J.0 [**Computer Applications**]: General; H.3.4 [**Information Storage and Retrieval**]: Systems and software—*Performance evaluation (efficiency and effectiveness)*

General Terms

User evaluation, Interaction, Patent Retrieval, PatOlympics

1. INTRODUCTION

Evaluating patent retrieval tools is problematic due to several factors. First, the users: experienced professionals whose job it is to use search tools to find relevant documents for a specific request for information. Second, the large experience concentrated in the top 3 patent search services providers in developing tools to address the needs of these very special kinds of users and the lack of willingness of these commercial providers to participate in standardized evaluation campaigns.

As a result, such evaluation campaigns, organized in the context of NTCIR [2], CLEF [3] or TREC [1] have a limited impact on the patent professionals (although they do have an impact overall [4]). To some extent, this is a result of our (i.e. IR researchers) incapacity to properly communicate to them the results obtained in such campaigns. However often enough, when we do get a user to take a look at a newly developed academic system, we discover that a lot of the work is replicating available features in commercials systems to which the researchers have no access and of which they are therefore unaware.

In the absence of cooperation from the commercial providers, the organizers of the Information Retrieval Facility Symposium[1], in cooperation with the organizers of the CLEF-IP and TREC-CHEM campaigns, have decided to put together an event which will bring together, in very close interaction, the users and the IR researchers. As a consequence of the experience of CLEF-IP and TREC-CHEM, the PatOlympics consist of two *patSports*: *CrossLingual Retrieving* and *ChemAthlon*. As their names suggest, CrossLingual Retrieving targets those systems which claim to be able to answer queries in one language with documents in other languages. ChemAthlon, on the other hand. is where systems being able to index and search chemical compounds (and not only) compete. Here the data and queries are always in English. While the two patSports require different approaches to the data and to the requests for information, in terms of the organization of the event, they are very similar and in the rest of the paper I shall refer to PatOlympics in general, and only refer to one of the patSports if something is really particular to it.

This paper is organized as follows: the conceptual and implementation design of the PatOlympics is described in Section 2. It then goes into the details of the logical back-end implementation in Section 3.1. It returns to the surface again in Section 3.2 with the description of the front-end interface. After all the details of the implementation, in Section 4 it describes the experience accumulated over the past two years in using it. Finally, conclusions and future work are described in Section 5

2. DESIGN

The idea of an event like the PatOlympics came from the organisers of the IRF Symposium in 2010 and the design, organization and implemention of the event was assigned to the author. The request was to have a half-day event, in the form of a competitive demo, where the two target groups of the Symposium (IR researchers and IP professionals) would be interacting on the basis of very specific tasks. The idea was to let the professional IP searchers test different participating systems on a particular request for information.

*The author is partially supported by the PROMISE NoE, funded under FP7 Grant no. 258191

DESIRE'11, October 28, 2011, Glasgow, Scotland, UK.

[1]http://www.irfs.at

The objective was to stimulate conversation between the two groups, to let IR researchers learn what is already available to the professional users and what still remains as a pain point in their day-to-day activities.

The requirement to make the event as a competition, in order to stimulate interest, lead to constraints in its design: we needed to make sure that whatever score was to be computed was comparable between participants. We shall describe in detail all the constraints and the solutions founds to match them in what follows. Before that, let us introduce some terms:

referee : an IP expert testing the systems in the PatOlympics with a specific request for information. A person can be a referee for one or for both patSports. For each of the patSports, the referee has one and only one request for information. Topics are not shared between referees.

team : generally an IR research group having developed a patent retrieval system and participating in either CrossLingual Retrieving or ChemAthlon.

participant : either a referee or a team.

Now let us return to the requirements that the PatOlympics must satisfy. The stem from the two different aspects of the event: the competition side and the interactive side.

Requirements stemming from the competitive nature of the event:

C1 *all teams in a patSport must work with the same topics.* In our context, this means that every team must interact with every referee, for every patSport it participates in.

C2 *all teams must have access to the same initial datasets.* The competition here is on IR methods, not on data availability. Therefore, the organizers make available to the participants, a few months in advance of the event, two data collections (one for each patSport). These data collections are the same as those given to CLEF-IP and TREC-CHEM participants in the same year.

C3 *all scores must be computed based on the same set of relevant documents for each topic.* We use a special form of pooling which makes sure that, by the end of the event, all scores are computed based on the same set of relevant documents.

C4 *all teams must be allowed to work on any particular topic exactly the same amount of time.* This is easily satisfied by only allowing each team to send in candidate relevant documents for a specific topic only during the round when they are scheduled to work with the referee owning that topic.

Requirements stemming from the interactive and cross-domain nature of the event:

I1 the interaction between a referee and a team must be of at least 20 minutes, in order for the referee to understand the system and be able to use it to find relevant documents

I2 the event cannot last for more than three hours in total, because both the referees and the teams would be exhausted.

I3 the scores must be updated continuously and displayed prominently in order to maintain the interest of the non-participating audience in the event.

I4 the score must be easily understandable by anyone in the room.

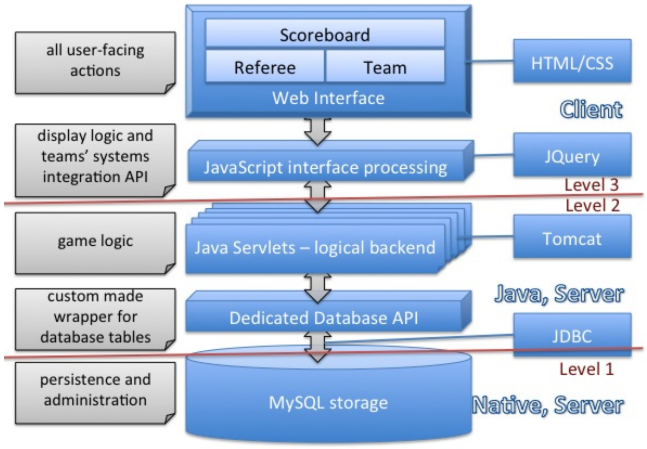

Figure 1: PatOlympics Application Architecture

I5 it must be easy for participants to submit candidate results to the system directly from their own prototype.

As a consequence of I3 and particularly of I4, we have not used Mean Average Precision or Normalized Discounted Cumulative Gain, but rather a simple Precision at 200 measure, even if the former are generally perceived as better IR evaluation metrics. Furthermore, we have displayed the Precision at 200 as simply "Number of relevant documents found" by each system, as each system was only allowed to send in up to 200 documents.

The combination of C1 with I1 and I2 results in the organization of the event in a series of up to 5-6 rounds, each of about 25 minutes, in which each referee sits at the table of a team, presents with his or her request for information and works together to find the answers. In the next Section we will describe how these rounds are generated.

There was also another requirement which does not fit into the above two categories: because the referees are business people, there is little to no time prior to the event to generate relevance judgements for the request for information that they have. To satisfy C3, the PatOlympics system must therefore allow the referees to indicate what they found to be relevant during the event itself, and separately from what the teams send in as their candidate answers (although they often overlap significantly).

3. THE PATOLYMPICS APPLICATION

The PatOlympics Application consists of three main layers (Figure 1):

Layer 1 is a native storage server in the form of a relational database (MySQL). It is used for persistence as well as for administration purposes, as everything above this layer takes its information from here.

Layer 2 is a web application hosted by a Tomcat Server. The web application consists of two packages. First, a set of wrappers for the tables in Layer 1 makes it easy for the programmer to access the relation database. Second, a collection of 28 servlets forms the main logical structure of the application. This set of servlets controls everything from user authentication and score calculations to the ticker message displayed on the score board. Tomcat exposes a subset of these servlets as webservices to be used by teams who wish to connect directly to the server.

Layer 3 is interpreted at the client side and consists

Figure 2: An example of optimal rounds assignment

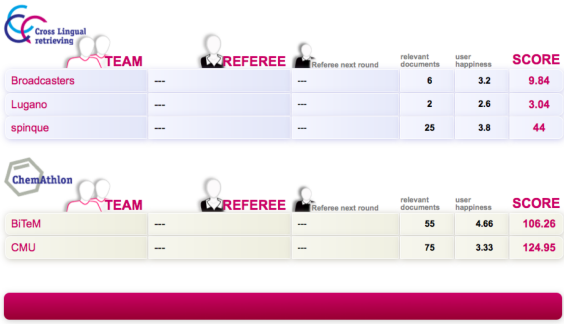

Figure 3: The final 2011 scoreboard

mostly of JavaScript to access the services of Layer 2 and HTML/CSS to display the pages in a browser.

3.1 Back-end

As previously mentioned, the back-end is the part which does all the application logic. The main part of this logic is the score computation. The score of each team is computed as a combination of two factors: the number of relevant documents found for each topic and an arbitrary "user satisfaction" score given by each referee at the end of the round, as a value between 1 (unsatisfactory) and 5 (very good). The precise formula for computing this score is: $finalScore = nRelRet \cdot (1 + 0.2 \cdot avgUseHap)$, where $nRelRet$ is the total number of relevant documents returned by the team across all topics of the patSport it is competing in, and $avgUseHap$ is the average user happiness scores given to this team by all the referees with which it has interacted. The formula takes into account the performance of the system in terms of retrieving relevant documents, but gives the system a small boost, proportional to the average user satisfaction score.

This score is recomputed every time one of 4 actions is taken: the referee adds or removes from the list of relevant documents one document; the team adds or removes an element from its list of candidate relevant documents. In each of these cases, the servlet managing this interaction triggers a score update and the new score is immediately stored in the database, from where it is then served to the interface to be displayed on the scoreboard.

Rounds generation.

Generating correct and efficient rounds is one of the main aspects of PatOlympics application. The problem is complicated by the fact that each team can participate in one or both patSports, just as well as each referee can have one or two topics (i.e. act as referee for only one or for both of the patSports). The constraints in this problem are:

1: each participant can only address one topic per round

2: each team can only be assigned a specific topic once in the entire game. But note that a team and a referee may meet a second time if both of them are in both patSports.

We approach the problem with a very simple, greedy method. First, we generate a matrix, whose columns are the teams and rows are referees. We initialize each cell with the number of times that particular referee must meet with that particular team. This will be 0 if they are not participating in the same patSport, 1 if they have one patSport in common and 2 if they both participate in the two patSports. We then greedily generate rounds by going through the matrix and every time we find a non-zero cell we assign the pair (team, referee) indicate by that cell to the current round and set to zero all other elements on the line and column of this cell. This is repeated until the entire matrix is zero.

However, this greedy method is not guaranteed to generate optimal results. However, because it is so fast, we can explore the entire space of possible assignments in less than a minute on a standard laptop. The method is therefore repeated as needed and the assignment with the minimal number of rounds is taken.

An example of a round assignment, for the participants of the 2011 event, is shown in Figure 2. As can be seen, most of the rounds are fully occupied, no team sees the same referee twice in the same patSport, and no referee is present twice in the same column (round).

3.2 Front-end

The most visible part of the front-end is the Scoreboard (Figure 3). During the event, this is displayed for all attendees of the symposium to see. Because it is just a web page, the Scoreboard can be displayed on multiple screens throughout the location of the IRF Symposium. Each interested attendee can also simply visit the PatOlympics website to see the latest scores of their favourite team.

The most important feature of the front-end is however the Referee interface. This is where the referees can indicate which documents are relevant for their request for information. They can do this at any time: before, during, as well as after the event.Figure 4 shows such an interface for the user Teresa (Loughbrough). We can see that this user is participating in both patSports, by the fact that two topics are available to her. Currently, she is working on the ChemAthlon one, as indicated by the red colour of the button at the top, as well as by the description and title of the topic displayed to her. Right under the description of the topic, in this example just a place-holder, she has the possibility to input documents relevant to it. The documents are identified by their UCID (the unique identifier of a patent document, formed by concatenating the country code, the patent number and the kind code). On the right hand side, we can see the list of patent documents already identified as being relevant to this ChemAthlon topic. For an easy identification, the list also displays the title of the patent. The referee can sort this list on each of the three columns (document ID, title, or time of addition to the list). She can also change her mind and remove a particular document by clicking on the red 'X' button on each row.

As mentioned previously, the interaction of the referee with the system, via this interface, triggers score updates for all participants in the current patSport. Basically, the user is dynamically modifying the qrels and therefore all

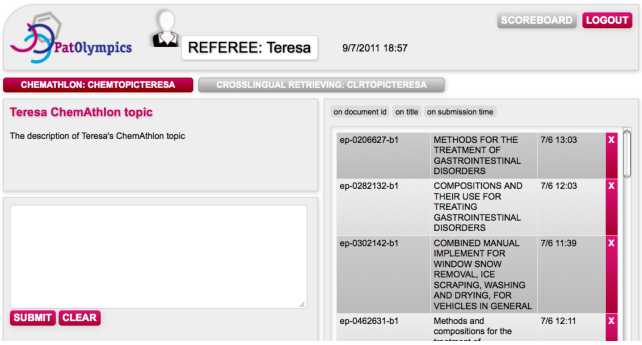

Figure 4: The Referee Interface

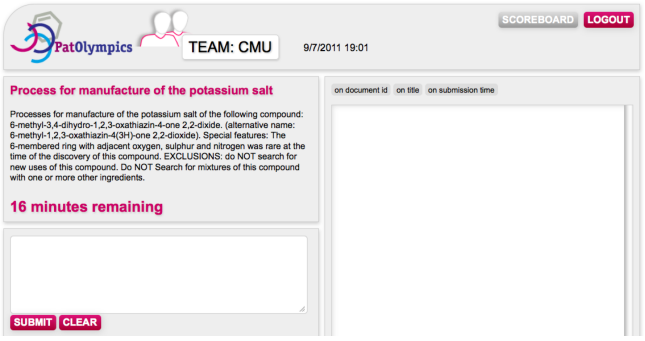

Figure 5: The Team Interface

participants' score are affected. This results in a very interesting experience, as the scoreboard will show updates even for teams which, due to the scheduling of the rounds, are not in the current round. This can very well happen if a team has used all its quota of 200 document candidates, and the referee identifies one of the as relevant at a later stage, while looking at the results of another system. This is the special pooling we mentioned in Section 2 as requirement C3.

The teams also have the possibility to use a web interface. However, as organizers, we encourage them to connect directly to the server using the API made available to them. This would not distract the user (be that the referee or the team itself) from the task at hand by requiring a constant shift between windows on the screen. Nevertheless, the option of using a web interface is useful as a back-up measure. In essence, this interface is the same as the one of the referee. The major difference comes as a consequence of requirement C4: a team is only allowed to submit results for a topic while the owner of the topic is assigned to work with them. This is why a timer is displayed just under the description of the topic they are supposed to be working on. When the time expires, the system automatically displays the next topic, or shows the time until a new topic will be available.

4. THE EXPERIENCE

2011 is the second year when the PatOlympics has been organized, and in both events the participants' feedback was very positive. It was recognized on both sides that this interaction is very useful. IR researchers learn a lot about how real patent searchers work, what are they looking for and how they approach a search problem. Patent searchers find out what can be done with today's technology and direct the researchers towards that of tomorrow.

Despite the inherent problems of a live demo (even more problematic when you have 6-7 demos running in parallel), the event has in both cases been seen as a success. The PatOlympics application has fulfilled its basic duties (computing, showing results), but further work on it is still needed.

In terms of the design of the event and of the scoring method, all of the participants agreed that this is a reasonable compromise between statistical validity of the results and practical possibilities and utility. A major problem in this type of exercise is the fact that it becomes extremely tiring after a relatively short time to dedicate the same intellectual effort to explain a topic as well as to understand and use a new system. Particularly in 2011, the rounds were very tight and some of the referees had to go constantly from team to team, just as some teams (the ones participating in the CrossLingual Retrieving) were subjected to a new topic at every round, without any breaks.

5. CONCLUSIONS AND FUTURE WORK

The PatOlympics is a well known and desired features of the IRF Symposium. As organizers, we are deeply grateful to both the referees and the teams participating. They are the ones really making the event what it is and it is here that organizing such an event presents its greatest challenge. For it is not easy to find any other event which brings together both IP professionals and IR researchers. Their simultaneous presence at the same place is vital. An IP expert finds an IR conference perhaps too technical, but most certainly focused on other aspects of the science behind the systems than what is really interesting for her. At the same time, and IR researcher finds an IP conference too expensive and unjustifiable in light of the lack of proper peer review. We must therefore strive to continue organizing this event and look to the two communities to acknowledge their mutually beneficial interaction.

In terms of the PatOlympics infrastructure, it requires further development to better handle unforeseen events during the games. For instance, there is currently no way to insert a break in the running of the games. Once started, the rounds run until their scheduled termination. Still, the design is competitive as it is, and it has never crashed neither during tests nor, more importantly, during the event.

Acknowledgments I thank Henk Tomas, Monika Hanelt, Tony Trippe, Teresa Loughrough, Pierre Buffet, Stephen Adams, Jane List, Maike Houtrouw, and Gerard Ypma for volunteering their time to prepare the topics and work with the participants. The first prototype of the software was developed with Binh Thanh Nguyen. Finally, the IRF management deserves the credit for the vision of the PatOlympics.

6. REFERENCES

[1] M. Lupu, J. Huang, J. Zhu, and J. Tait. TREC Chemical Information Retrieval - An Evaluation Effort for Chemical IR Systems. *WPI Journal*, 2011.

[2] H. Nanba, A. Fujii, M. Iwayama, and T. Hashimoto. Overview of the Patent Retrieval Task at the NTCIR-8 Workshop. In *Proc. of NTCIR-8*, 2010.

[3] F. Piroi and J. Tait. Clef-ip 2010: Retrieval experiments in the intellectual property domain. In *Proc. of CLEF*, 2010.

[4] B. Rowe, D. Wood, A. Link, and D. Simoni. Economic Impact Assessment of NIST's Text REtrieval Conference (TREC) Program. Technical report, National Institute of Standards and Technology, 2010.