Little Search Game: Term Network Acquisition via a Human Computation Game

Jakub Šimko
Institute of Informatics and Software Engineering,
Faculty of Informatics and Information Technologies,
Slovak University of Technology,
Bratislava, Slovak Republic
jsimko@fiit.stuba.sk

Michal Tvarožek
Institute of Informatics and Software Engineering,
Faculty of Informatics and Information Technologies,
Slovak University of Technology,
Bratislava, Slovak Republic
tvarozek@fiit.stuba.sk

Mária Bieliková
Institute of Informatics and Software Engineering,
Faculty of Informatics and Information Technologies,
Slovak University of Technology,
Bratislava, Slovak Republic
bielik@fiit.stuba.sk

ABSTRACT

Semantic structures, ranging from ontologies to flat folksonomies, are widely used on the Web despite the fact that their creation in sufficient quality is often a costly task. We propose a new approach for acquiring a lightweight network of related terms via the *Little Search Game* – a competitive browser game in search query formulation. The format of game queries forces players to express their perception of term relatedness. The term network is aggregated using "votes" from multiple players playing the same problem instance. We show that nearly 91% of the relationships produced by *Little Search Game* are correct and also elaborate on the game's unique ability to discover term relations, that are otherwise hidden to typical corpora mining methods.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Content Analysis and Indexing; K.8 [**Personal Computing**]: Games

General Terms

Design, Experimentation

Keywords

human computing, games with a purpose, term network

1. INTRODUCTION & RELATED WORK

Semantic metadata and structures are a vital part of many web-based applications like faceted browsers [12] or (personalized) e-learning systems [3], which utilize ontologies and properly annotated resources to enable filtering of large corpora and recommendation of relevant content. Even more lightweight structures like folksonomies are sufficient for use in the Semantic Web, for example to guide searchers during exploratory search tasks [8] by expanding their queries [1, 4], navigating and visualizing the information space [11] or structuring the search history for revisitation purposes [14].

Effective creation of "high quality" semantic structures is often a non-trivial task which scales badly with size and complexity. Perhaps the best results in terms of high quality and trustworthiness of semantic structures can be achieved by using manual creation by human experts. If a project is given enough time, like in the case of the Cyc ontology [7] or enough participants (DBPedia) it can eventually grow in quantity, but generally, human experts are sparse and costly. On the other hand, quantity can be easily provided by automated approaches, such as mining text corpora using latent semantic analysis [9]. The drawbacks are often in the lacking quality of automatically acquired semantics (e.g., due to problems with natural language processing) [15]. To achieve quantity, "indirect" crowdsourcing for folksonomies may be an alternative (e.g., Delicious, which exploits coocurrence of terms in user created resource annotations), but it operates over too general terms.

However, the crowdsourcing approach used in *games with a purpose* (GWAP) is potentially capable of overcoming the aforementioned problems. It emerged as an alternative approach to solving computational problems, hard or impossible to be solved by machine computation (which includes acquiring semantic structures), via aggregation of knowledge provided by many non-expert users (e.g., for image annotation) [10]. GWAPs transform problems into games that motivate players to solve them via fun and thus eliminate the need to pay them. As many game instances can be played simultaneously, they are suitable for larger scale problems divisible into smaller tasks. Compared to other crowdsourcing techniques, the knowledge gained in GWAPs is not just a by-product of another user activity (e.g., annotating web resources for personal use), but the *primary* objective, so their design is tuned to maximize that ability (e.g., overcoming the "too general terms problem" as proposed in our method described in this paper).

The use of games with a purpose for the acquisition of semantics and semantic structures was pioneered by Luis von Ahn in his *ESP Game* [13], where two players try to agree on a single word describing a given image, thus effectively tagging it (which is normally unsolvable by a machine). Ahn

HT'11, June 6–9, 2011, Eindhoven, The Netherlands.

also shows the key aspects of GWAPs: human effort paid for by the non-monetary value of entertainment, possible massive parallelization by multiple players and self-correction of the game results by collaboration and agreement of the individual player outputs [13].

Several games were specifically designed for the Semantic Web environment. *Verbosity* is a two player word guessing game that collects facts in a form suitable for ontology construction – players have to provide clues using predefined sentence templates that refer to named relationships [13]. The drawback of *Verbosity* is that most often not all possible relationships can be retrieved (even unnamed ones) . Other addressed tasks related to Semantic Web include linking (the purpose of *OntoPronto*) and ontology alignment (in *SpotTheLink*) [10]. Additional games devised to create resource metadata include *Peekaboom* [13] (acquisition of object positions in images) or Phrase detectives [5] (identification of relations between words and phrases in text).

In this paper, we present the *Little Search Game* – an implementation of GWAP principles with the purpose of creating a network of unnamed term relationships (as seen in Figure 1) with these advantages: (i) the relationships (though unnamed) cover all types instead of being of a predefined set, (ii) the network contains more specific terms than in regular folksonomies as we decide for which terms relationships are added, (iii) the network includes relationships, that are not supported by real web co-occurrence of terms but are semantically sound (i.e., relations that cannot be discovered by automated corpora analysis).

2. LITTLE SEARCH GAME

We devised the *Little Search Game* to create lightweight, folksonomy-like term relationship networks. It is a search-query formulation game where the player's task is to reduce the number of results returned by a search engine to an initial query term, by guessing the best negative search terms to be attached to the query that reduce the number of results from the original result set. In order to succeed, players must enter negative terms with high web co-occurrence with the initial term, which from their point of view means: "somehow related to it". These "opinions" are then aggregated into the collaboratively created term relationship network.

Playing the game is motivated via competition among players and via the challenge of overcoming oneself with better a score. The game rules are described in the following scenario (Figure 2 shows the corresponding game screen):

1. First, players are given an initial query that consists of a single *task term* (e.g., "star") which returns a certain number of search results (e.g., 2 billion).

2. Next, players guess *negative terms* that expand the original query (e.g., "star –movie –wars –death") so that it returns fewer results (around 400 million in the example in Figure 2). The query format utilizes the "–" directive instructing search engines to reduce the original result set via terms decorated with it.

3. The lower the final number of results, the better is the resulting score. To succeed, one must enter "proper" negative terms, i.e. those with high co-occurrence with the initial term. Players interpret this as a "relationship" between terms and try to enter terms they consider to be mostly related to the original term. Players can perform multiple attempts to improve their score.

The creation of the term relationship network is a player agreement process where each original triplet of (*player*, *task term*, *negative term*) is considered a *vote*. If at least N votes with the same *task term – negative term* combination exist (i.e., N different players consider those words related) a new oriented relationship (from task term to negative term) is created in the network (see Figure 1). In our experiments, we chose $N = 5$, which proved to be sufficient.

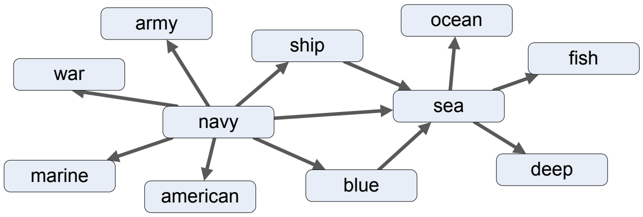


Figure 1: Subset of the created term network.

We deployed the game as a browser applet called the *Little Google Game*[1] (since we employed Google as the underlying search engine). We deployed the game in several modes to gather user feedback and data for term network creation. In tournament mode, we used the game to stage a tournament among contestants and awarded prizes to the best players; in showcase mode the game was available online to the public and during a showcase exhibition in a shopping mall.

So far, the game was played by about 300 players with 3,800 played games and 27,200 submitted queries. Players guessed over 3,200 negative terms. The total amount of task terms used in the game so far was 40. The tasks terms were chosen randomly, however task selection can be driven by external demands (e.g., focused network extension with specific terms). The resulting network contains 400 nodes and 560 edges. The distribution of relationships per task term differs between game modes (tournament tasks were played much more often than tasks used during the showcase). By imposing an additional log analysis rule, that only 10 strongest relationships per task term were evaluated, the resulting network shrunk to 183 nodes and 220 edges.

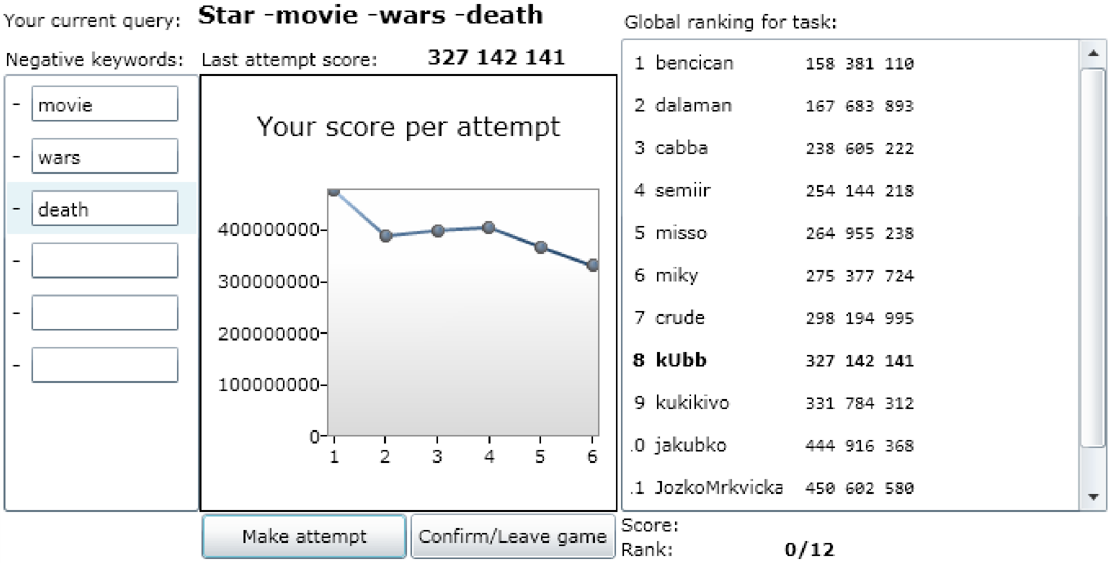
3. TERM NETWORK VALIDATION

We validated the following hypothesis: *Every (oriented) edge in the term network created by Little Search Game reflects a real semantic relationship of the source term with the target term.* To do so, we conducted a survey evaluating the soundness of a subset of the created term network.

Participants. The survey was conducted with 18 participants – adults aged between 18 and 35 with high school, undergraduate or graduate education in various professions. All of them spoke at least intermediate level English and were not aware of the background of this survey.

Data. We randomly chose 12 relationships from the *Little Search Game* term network to be evaluated. To create some "noise", so that participants would not realize they were expected to mark each relationship with a positive vote, we created 8 more random term pairs and shuffled them into the original 12 to create a list of 20 ordered term pairs.

[1] mirai.fiit.stuba.sk/LittleGoogleGame



Your current query: **Star -movie -wars -death**

Negative keywords:
- movie
- wars
- death

Last attempt score: **327 142 141**

Global ranking for task:

1	bencican	158 381 110
2	dalaman	167 683 893
3	cabba	238 605 222
4	semiir	254 144 218
5	misso	264 955 238
6	miky	275 377 724
7	crude	298 194 995
8	**kUbb**	327 142 141
9	kukikivo	331 784 312
.0	jakubko	444 916 368
.1	JozkoMrkvicka	450 602 580

Score:
Rank: **0/12**

Figure 2: The game interface shows the current negative query terms (left), the history of attempts (center) and the global player ranking for the given task (right).

Task. Participants were presented with a list of ordered term pairs (referred to as term A for source terms and B for target terms), with the task: *"Do you consider the term B as being related to the term A (in other words: would you include the term B in the top 10 most related terms for term A)? (1 – Definitely irrelevant, 2 – More likely irrelevant, 3 – More likely relevant, 4 – Definitely relevant, 5 – Unsure)."* We indirectly stressed the importance of evaluating a one-way relationship from A to B, since our term network is an oriented graph.

Environment. Participants were not put under any time stress. They completed the questionnaires either in paper or electronic version. No further suggestions on how to populate the forms were given to them.

Results. We evaluated the answers as follows (see Table 1):

1. We counted the selected options for each pair. Counts can be seen in columns *1, 2, 3, 4* and *unsure*.
2. We counted votes *for* and *against* the relevance of particular term pairs. Options *1* and *4* were counted twice since participants were more certain when they used them and did not know how the survey would be interpreted. *Unsure* votes were excluded.
3. We computed whether the participants as a group rejected or admitted the relevance of term pairs based on vote counts. *Mass opinion on relevance* was set to *yes* or *no* if one of the respective weighted vote counts was at least twice that strong than the other one. The rest of the pairs was set as *controversial* and the pair was removed from further evaluation.
4. Finally we checked how many of the term pairs in the network were marked relevant by voters (*success* column). We did not evaluate the "noise" term pairs.

The strength of the pairs within the term network is shown in the column *weight in the network* (ω_p/ω_t). The results were more than encouraging – almost **91%** of the term relationships in the created term network were judged **correct**.

4. HIDDEN RELATIONSHIPS

The interpretation of frequent co-occurrence of two terms on the Web as the existence of a semantic relationship between them is not accurate – not all related terms frequently appear together. Meanwhile there are term pairs with significant co-occurrence but no real semantic link between them. This generates noise, which prevents us from using statistical methods for discovery of all relevant term relationships.

Little Search Game enables us to discover these hidden relationships: players sometimes use negative terms, which they believe to be related to the task word and should decrease the number of search results, but have in fact almost no effect on the result set. Even if they later abandon these terms, their single initial use is sufficient for term network extraction. In order to discover the significance of this approach, we conducted a statistical experiment to estimate the number of relevant term pairs that are impossible to discover due to the aforementioned noise.

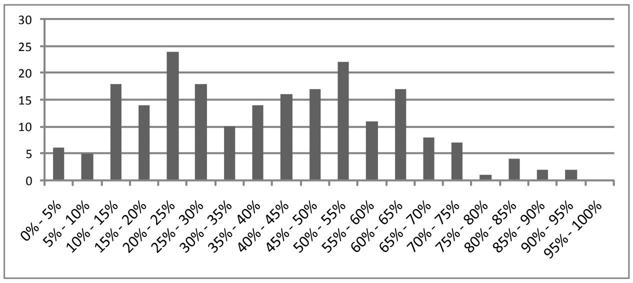


Figure 3: Distribution of term network relations (oriented term pairs) by the real co-occurrence of the paired terms from the source term's standpoint.

Table 1: Result sheet of the Little Search Game term network validation survey. 12 term pairs from the term network were mixed with 8 random pairs. Participants voted for options 1 to 4 (1 – definitely not related, 4 – definitely related) to evaluate, whether the term B is related to term A.

Term A		Term B	Survey answer counts					Against	For	Mass opinion on relevance	Weight in the term network	Success
			1	2	3	4	Unsure					
brain	-	game	2	4	7	5	0	8	17	Yes	10/173	Yes
firefighter	-	white	13	4	0	1	0	30	2	No	0/0	N/A
cellular	-	biology	0	1	5	12	0	1	29	Yes	33/514	Yes
einstein	-	physics	0	0	0	18	0	0	36	Yes	16/187	Yes
remove	-	island	10	6	2	0	0	26	2	No	0/0	N/A
einstein	-	iq	0	2	6	10	0	2	26	Yes	8/187	Yes
navy	-	marine	0	2	1	15	0	2	31	Yes	12/146	Yes
desert	-	glass	7	8	2	1	0	22	4	No	0/0	N/A
forest	-	movie	10	2	4	2	0	22	8	No	15/604	No
sea	-	deep	0	0	8	10	0	0	28	Yes	9/168	Yes
hydrogen	-	browser	15	2	0	0	1	32	0	No	0/0	N/A
star	-	sky	0	0	4	14	0	0	32	Yes	9/154	Yes
brain	-	tumor	1	2	6	9	0	4	24	Yes	7/173	Yes
cellular	-	network	1	4	6	6	1	6	18	Yes	19/514	Yes
einstein	-	nobel	0	1	8	9	0	1	26	Yes	7/187	Yes
cellular	-	automat	5	4	3	3	3	14	9	Controversial	20/514	N/A
church	-	food	9	6	2	1	0	24	4	No	0/0	N/A
fast	-	radiation	5	9	1	3	0	19	7	No	0/0	N/A
heat	-	war	4	6	5	3	0	14	11	Controversial	0/0	N/A
theatre	-	technology	9	6	0	3	0	24	6	No	0/0	N/A

Figure 3 shows the distribution of term pairs by their co-occurrence rate r_s. Term pairs are taken from the reduced term network (only top 10 relationships for each source term are included) created by *Little Search Game*.

We computed the co-occurrence rate of term pairs as follows: For each relationship in the term network (during the experiment, we had 216) we conducted 3 search queries to acquire 3 values p_s, p_t, i, where p_s and p_t are numbers of results returned by a web search engine for the source and target terms separately, and i is the number of returned results for the combined query *"[source term] AND [target term]"*, effectively giving us the size of the intersection of result sets of individual terms. We computed values r_s and r_t defined as size ratios of the intersection result set and term result sets:

$$r_s = \frac{i}{p_s}, r_t = \frac{i}{p_t}$$

The r_s value is the co-occurrence rate with the target term from the source term's perspective.

As shown in Figure 3, the term network contains term pairs with varying real co-occurrences on the Web. The ones with high rates can be discovered statistically, simply by querying for all possible combinations of terms and collecting relatively high r_s and r_t values. To discover, which rates are sufficient for a relationship, the noise distribution must be determined, i.e. the distribution of rates r_s and r_t for non-related term pairs randomly taken from the language corpus.

Determining the noise distribution is complicated due to varying term frequencies in the language itself. One extreme are stop words, which are present in practically every document. Their inclusion into the set of words from which we pick random pairs would cause a high level of noise that would prevent proper statistical relationship discovery. The other extreme are specialized words with very low frequencies (most of the language corpus). Selecting pairs out of them would suggest, that there is no noise at all, since they have virtually no co-occurrence.

Therefore, we conducted the noise computation for three different-sized word corpora (800, 5,000 and 50,000 terms), populated with most frequently used words in the English language, excluding stop words. For each corpus, we randomly picked 200 term pairs, computed their values r_s and r_t, computed the value $r; r = r_s/r_t$ as their average and plotted its distribution, shown in Figure 4.

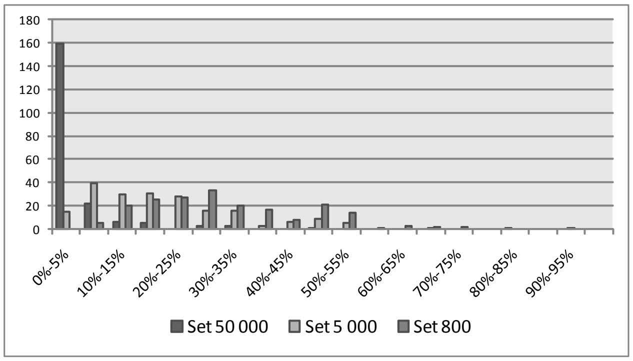


Figure 4: Term co-occurrence noise shows distribution of the co-occurrence rates for random term pairs. Computed for three different sized term sets.

As expected, the smallest set yielded the highest noise, significant already at 50% co-occurrence ratios. On the other hand, in the largest corpus, the noise level is minimal with 95% of term pairs with r value less than 5%.

The interesting noise levels were in the medium sized corpus of 5,000 terms, where noise starts to be significant at

35% of the co-occurrence ratios. Almost all of terms in the *Little Search Game* term network can be found in this corpus. Assuming that the noise significance starts at 35% rate of co-occurrence, the number of relationships in the term network with rates below that (as observed in Figure 3) is almost 40%, which means that 40% of the relationships discovered by *Little Search Game* can not be discovered by statistical co-occurrence analysis. This underscores the invaluable power of human computation and importance of our approach as a tool for releasing that power.

We used the Bing WSDL service in this experiment instead of Google, which was used in the game itself, because Google "spoils" the result sets with a significant number of additional results as it tries to improve search results. While that may be helpful when searching the Web, in our experiment it caused an unwanted bias (e.g., sometimes the result set expected to be the intersection was even larger than one of the original sets). Since Bing search followed boolean logic in queries unlike Google, we utilized it this experiment.

5. CONCLUSIONS AND FUTURE WORK

We presented *Little Search Game* – a game with a purpose for the acquisition of unnamed term relationship networks, which are useful in many Semantic Web applications. In this search query formulation game, players are obliged to provide sets of terms related to a starting term. Our evaluation has shown that 91% of relationships in the network are correct.

Our contribution is that the created term network:

- is more specific than common folksonomies and its growth can be intentionally directed,
- has no limits in terms of relationship types,
- includes a significant number of relations that cannot be discovered by purely statistical corpora analysis.

We see several possibilities for future work. The game itself could be improved to make its purpose and rules better understandable by players, since this is the main issue preventing its viral spreading on the Web. Another open problem is the refinement of the term network structure: the labeling of relations, switching from terms to concepts. Lastly, one could explore the possibilities of creating term relationships "on demand", since the game can use arbitrary terms as gameplay tasks. One possible example of such game usage is the conjunction with our adaptive proxy-server for web search [6, 2] where term relationships could be used to improve grouping of user profiles or to expand popular queries.

6. ACKNOWLEDGMENT

This work was partially supported by the grants VG1/0675/11/2011, KEGA 028-025STU-4/2010, APVV-0208-10 and it is the partial result of the Research & Development Operational Programme for the project Research of methods for acquisition, analysis and personalized conveying of information and knowledge, ITMS 26240220039, co-funded by the ERDF.

7. REFERENCES

[1] J. Bai, D. Song, P. Bruza, J. Nie, and G. Cao. Query expansion using term relationships in language models for information retrieval. In *Proc. of the 14th ACM int. conf. on Information and knowledge management*, volume 05, pages 688–695. ACM, 2005.

[2] M. Barla. Towards social-based user modeling and personalization. *Information Sciences and Technologies Bulletin of the ACM Slovakia*, 3(1):52–60, 2011.

[3] M. Barla, M. Bieliková, A. B. Ezzeddinne, T. Kramár, M. Simko, and O. Vozár. On the impact of adaptive test question selection for learning efficiency. *Computers and Education*, 55(2):846 – 857, 2010.

[4] J. Bhogal, A. Macfarlane, and P. Smith. A review of ontology based query expansion. *Inf. Process. Manage.*, 43:866–886, July 2007.

[5] J. Chamberlain and M. Poesio. A Demonstration Of Human Computation Using The Phrase Detectives Annotation Game . *Discourse*, pages 23–24, 2009.

[6] T. Kramár, M. Barla, and M. Bieliková. Disambiguating search by leveraging the social network context based on the stream of user's activity. In *UMAP '10: Proc. of the 18th Int. Conf. on User Modeling, Adaptation, and Personalization*, pages 387–392, Hawaii, HI, USA, 2010. Springer.

[7] D. B. Lenat. Cyc: a large-scale investment in knowledge infrastructure. *Commun. ACM*, 38(11):33–38, 1995.

[8] G. Marchionini. From finding to understanding. *Communications of the ACM*, 49(4):41–46, 2006.

[9] L. A. F. Park and K. Ramamohanarao. An analysis of latent semantic term self-correlation. *ACM Trans. Inf. Syst.*, 27:8:1–8:35, March 2009.

[10] K. Siorpaes and M. Hepp. Games with a Purpose for the Semantic Web. *IEEE Intelligent Systems*, 23(3):50–60, May 2008.

[11] R. Stewart, G. Scott, and V. Zelevinsky. Idea navigation: structured browsing for unstructured text. In *CHI '08: Proc. of the 26th SIGCHI conf. on Human factors in computing systems*, pages 1789–1792, New York, NY, USA, 2008. ACM.

[12] M. Tvarožek and M. Bieliková. Generating exploratory search interfaces for the semantic web. In *Human-Computer Interaction*, volume 332 of *IFIP Advances in Information and Communication Technology*, pages 175–186. Springer Boston, 2010.

[13] L. von Ahn and L. Dabbish. Designing games with a purpose. *Communications of the ACM*, 51(8), 2008.

[14] J. Šimko, M. Tvarožek, and M. Bieliková. Semantic History Map : Graphs Aiding Web Revisitation Support. *Proc. of theWorkshops on Database and Expert Systems Applications*, pages 206–210, 2010.

[15] T. Wang, D. Maynard, W. Peters, K. Bontcheva, and H. Cunningham. Extracting a domain ontology from linguistic resource based on relatedness measurements. In *WI '05: Proc. of the 2005 IEEE/WIC/ACM Int. Conf. on Web Intelligence*, pages 345–351, Washington, DC, USA, 2005. IEEE CS.