Augmenting Human Computed Lightweight Semantics

Jakub Šimko*
Institute of Informatics and Software Engineering
Faculty of Informatics and Information Technologies
Slovak University of Technology in Bratislava
Ilkovičova 3, 842 16 Bratislava, Slovakia
jsimko@fiit.stuba.sk

Abstract

Semantic web has to overcome several challenges ranging from web resource annotation to domain modelling. Acquiring general or domain knowledge ontologies is done by automated approaches with only limited sucess or is left to costly human experts. Games with a Purpose offer an opportunity to employ broader crowd of laics into these tasks by transforming tasks to the appealing fun. We build upon our previous research of the *Little Search Game*, a search query formulation game for acquiring lightweight network of unnamed term relationships. We analyze these relationships for types and argue for suitability of this network to provide non-taxonomic relationships, so much needed in existing knowledge bases.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; K.8.m [**Personal Computing**]: Games

Keywords

Games with a purpose, semantics, web, lightweight, human computation, web search

1. Introduction and Related Work

The semantic web and its principles promise enhancements of our web experience in many ways. Philosophically, it offers automated processing of the web content on the conceptual level, which corresponds to the way of humans. Rich metadata enables more meaningful web search results to be retrieved. Semantically connected resources and concepts are employed in visualization of the domain helping to better understand it and aid during learning tasks [5]. Sophisticated reasoning over facts on the web can be used for information aggregation out of resources, which would normally be too heterogeneous for such task to be done automatically. Many approaches such as recommenders [3] or user modeling systems [1] also benefits from the structured semantics.

Provisions of the semantic web are evident but still, the domain ontologies and resource metadata lack scale and quality. Manual ontology engineering deliver high quality models, but it requires costly human effort. Automated approaches (e.g. natural language processing methods [2], particularly latent semantic analysis [4]) can reach sufficient scale easily and already brought some valuable results, however, the produced metadata need additional human validation. The crowdsourcing techniques are also capable of providing quantity and are especially useful for resource annotations (bookmarking portal *Delicious*). However, they have only limited capability in acquisition of complex semantic structures.

Many approaches focus on acquisition of more lightweight semantic structures, which in return for sacrificing some of the ontology's formal strength offer easier creation, while retaining usability for some problems (text similarity, content recommenders). These structures include taxonomies, crowdsourced folksonomies and other "loose-relationship" networks. Their entities usually do not hold any attributes and underlie under no constraints. However, these lightweight structures can be possibly promoted to ontologies, since they reflect hidden but still valid semantics.

Labeling those relationships means determining their type (e.g., hierarchical, compositional, interactional or some domain-specific type). While discovery of hierarchical relationship seems to provide high quality results even for automated approaches, exploring and consequently naming the non-taxonomic relationships is more challenging. Weichselbraun et al. use a finite set of possible relationship types within the domain and then matches them against set of unnamed relationships and text corpus of the domain [8]. Naturally, the input set of unnamed relationships determines, what we can expect as an output of the naming process (e.g., if the input set implicitly comprises only hierarchical relationships, we cannot expect other than "is-a" labels being assigned). In this paper, we argue that our approach of acquiring unnamed relationships (*Little Search Game*) has no such relationship type bounds and is suitable for enriching existing ontologies, especially with non-taxonomic relationships.

*Doctoral degree study programme in field Software Engineering. Supervisor: Professor Mária Bieliková, Institute of Informatics and Software Engineering, Faculty of Informatics and Information Technologies, STU in Bratislava. Work described in this paper was presented at the 7th Student Research Conference in Informatics and Information Technologies IIT.SRC 2011.

Šimko, J. Augmenting Human Computed Lightweight Semantics. Information Sciences and Technologies Bulletin of the ACM Slovakia, Special Section on Student Research in Informatics and Information Technologies, Vol. 3, No. 2 (2011) 116-118

2. Little Search Game

The analysis of promotion of lightweight semantics to richer structures is subject of this paper. In our previous research works [7], we introduced the *Little Search Game* (LSG), a human computation game for acquisition of a general term relationship network. In the game, players disclose relationships by formulating negative search queries (e.g. "sea –blue –fish –ocean"). We have validated the semantic soundness of the relationships within the network [7] and shown that the game can discover relationships that remain hidden to automated corpora mining methods[7]. The game belongs to the realm of games with a purpose, which are computer games that align solving of a human intelligence task, such as ontology building, into appealing game rules [6].

The hidden relationships, concept-based method of creation and overall "human origin" are the main arguments to use the *Little Search Game's* term network to enrich existing ontologies (especially those created with automated approaches, as they may utilize the hidden relationships). Our intents are mainly to introduce more non-taxonomic ontology relationships that are problematic to obtain. In that course we need to overcome two obstacles: promotion of the network terms to concepts and naming the network relationships. In this paper, we discuss the second issue.

We have conducted experiments to disclose how many of the relationships are present in the major knowledge base of Wikipedia using *Wikipedia Miner* tool and what kind of relationships are mostly present in the network, by evaluating them manually and also comparing them to the facts in the knowledge base of *ConceptNet*.

3. Experiments

We examined the term relationship network (LSG network) yet acquired to determine its potential for further upgrades to an ontology. In the first experiment (analogous to the "hidden relationship experiment" described in [7]) we questioned which of the LSG relationships can be considered hidden according to the human-created knowledge base - the Wikipedia. The second experiment was focused to determine the types of the relationships in our term network by comparing them to known relationships in the knowledge base of *ConceptNet*.

Our original "hidden relationship detection" experiment was done by comparing term (web) co-occurrence ratios of the LSG network relationships and a reference "noise" set, which consisted of random, nonsense term pairs. The co-occurrence of terms in each relationship (pair) was acquired via the search engine result set cardinalities for individual terms and their intersections. Relationships of the LSG network were considered hidden, if their co-occurrence ratios were equal or below the ratios of the non-sense pairs (i.e., they were undetectable by statistical analysis) [7].

The experiment described below was a modification of this approach. We just switched the co-occurrence ratio metric for the *Wikipedia Miner* concept relatedness. The *Wikipedia Miner* tool[1] computes semantic distance between the pairs of terms according to the linkage between so-named concepts in Wikipedia. By that, we examined how many of the LSG relationships are reflected in the

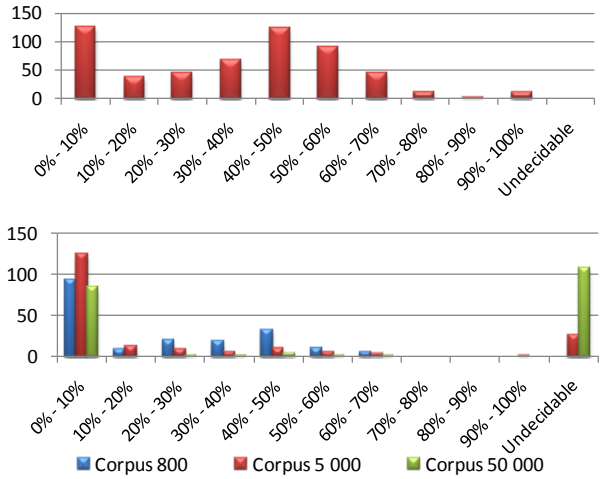

Figure 1: Upper: Relationships of the LSG network distributed by Wikipedia Miner relatedness. Lower: Non-sense relationships of the reference sets distributed by Wikipedia Miner relatedness.

Wikipedia (as the major knowledge base), since it is relevant to ask, how many new relationships the *Little Search Game* can offer to enrich the existing knowledge bases.

We used 400 LSG network relationships available at the time of the experiment. As reference sets, we used three sample sets of 200 nonsense relationships, comprising terms of three differently large corpora of most frequent words of the English language (800, 5 000 and 50 000 excluding stop words). Three sets were used due to expectation that different generality of terms render different noise levels (highest noise for smallest set). The *Wikipedia Miner* was queried for all term pairs of LSG network and reference sets (the relatedness is expressed by percentage). The evaluated relationships were grouped by 10% intervals. During the process, some of the relationships became unresolvable, because not all terms were present in the Wikipedia as concepts (Figure 1).

In the Figure 1 we can see a large portion (almost one quarter) of LSG network relationships having relatedness of their terms at near zero level. Second large portion (roughly another quarter) is being distributed in the 0%-50% relatedness interval, in which noise levels for smallest and middle reference sets are notable. This renders these relationships hidden, which means that it is relevant to do effort to incorporate them in the knowledge base such as Wikipedia's system of concepts. Unfortunately, experiment results may be biased in several ways. *Wikipedia Miner* was able to assign concepts to all given terms from LSG network, however it was unable to do so in case of some terms from reference set (especially for largest set, where terms were too specific to be contained within Wikipedia). The second issue may be wrong disambiguation of the terms in *Wikipedia Miner* (it works with the most frequent meanings) and thus pushing some of the LSG network relationships to zero relatedness.

In the second experiment we conducted, we queried the *ConceptNet*[2] (existing general knowledge base of common

[1] http://wikipedia-miner.sourceforge.net/

[2] http://csc.media.mit.edu/conceptnet

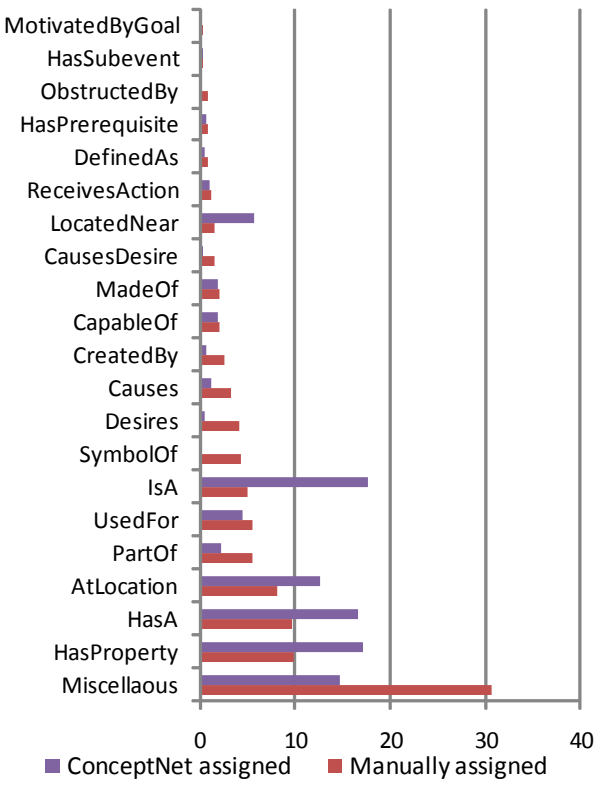

Figure 2: Relative distribution of relationship types in the manually evaluated set and ConceptNet.

facts) for types of relationships of the LSG network to show how many of them are really represented in such knowledge base and of what types they are. Additionally, we conducted a manual (two judge agreement) evaluation of all the relationships in the LSG network and assigned each of them the overall semantic soundness and one of the relationship types. This allowed us to discover more about the character of the LSG network.

We worked with 400 relationships created by *Little Search Game*, the relationships were sorted according to their strength based on how many votes they received during gameplay. Two independent judges manually evaluated each relationship in the LSG network. Both judges evaluated the soundness with one of the three values (sound, maybe sound and not sound), which after merging of both evaluation yielded five possible values. Judges also assigned one of the 23 relationship types (e.g., IsA, HasA, UsedFor, CapableOf) to each LSG relationship with possibility to assign a default unknown option, which was also imposed if the judges had not reached the agreement. The automatic retrieval of relationship types was done and all relationships were acquired.

The manual evaluation has shown that from 400 examined relationships 80% was semantically sound 8% rather controversial and 12% not sound (which is less than in the previous experiment [7], however, the "strongest" 100 relationships retained 93% soundness). The *ConceptNet* comprises only 164 (41%) out of 400 relationships we worked with. We consider it as strong argument for putting effort to enriching such knowledge bases.

The distribution of relationship types in the manually and *ConceptNet* annotated set differs (as shown on the Figure 2). First, high amount of the *ConceptNet* evaluated relationships is taxonomic (IsA). On the other hand, various dependencies (Desires, Causes...) almost absent in the *ConceptNet*. Generally, the *ConceptNet* relies roughly on 6 types of relationships, while the manually annotated set appears to be richer in types (for which we also argue to put effort to naming of the LSG relationships).

4. Conclusions

We analyzed the possibilities of enriching existing knowledge bases with triplets based on the unnamed term relationships produced by the *Little Search Game*, which we developed in our previous research. By examining Wikipedia and the ontology of *ConceptNet*, we have shown that current knowledge bases comprise only limited amount of all potential relationships and also that *Little Search Game* can provide more of them (of many different types), though they are not named yet.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0675/11.

References

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

[2] P. Buitelaar, P. Cimiano, a. Frank, M. Hartung, and S. Racioppa. Ontology-based information extraction and integration from heterogeneous data sources. *International Journal of Human-Computer Studies*, 66(11):759–788, Nov. 2008.

[3] M. Kompan, D. Zeleník, and M. Bieliková. Methods for personalized recommendation of nnewspaper articles. In *Znalosti 2011*, 2011. (In Slovak).

[4] 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.

[5] M. Tvarožek. Exploratory search in the adaptive social semantic web. *Information Sciences and Technologies Bulletin of the ACM Slovakia*, 3(1):42–51, 2011.

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

[7] J. Šimko, M. Tvarožek, and M. Bieliková. Little search game: Term network acquisition via a human computation game. In *Proceedings of the 22nd ACM Conference on Hypertext and Hypermedia*, pages 57–61, 2011.

[8] A. Weichselbraun, G. Wohlgenannt, and A. Scharl. Refining non-taxonomic relation labels with external structured data to support ontology learning. *Data & Knowledge Engineering*, 69(8):763–778, Aug. 2010.