Semantics Discovery via Human Computation Games

Jakub Šimko*
Slovak University of Technology, Slovak Republic

Michal Tvarožek
Slovak University of Technology, Slovak Republic

Mária Bieliková
Slovak University of Technology, Slovak Republic

ABSTRACT

The effective acquisition of (semantic) metadata is crucial for many present day applications. Games with a purpose address this issue by transforming computational problems into computer games. We present a novel approach to metadata acquisition via *Little Search Game* (LSG) – a competitive web search game, whose purpose is the creation of a term relationship network. From a player perspective, the goal is to reduce the number of search results returned for a given search term by adding negative search terms to a query. We describe specific aspects of the game's design, including player motivation and anti-cheating issues. We have performed a series of experiments with *Little Search Game*, acquired real-world player input, gathered qualitative feedback from the players, constructed and evaluated term relationship network from the game logs and examined the types of created relationships.

Keywords: Games with a purpose, human computation, crowdsourcing, metadata, semantics, search query, web search, term network

INTRODUCTION

Knowledge and semantics are needed both in quality and quantity. Many contemporary applications rely on (semantic) metadata in order to provide their intended functionality (Siorpaes & Simperl, 2010). Consequently, the creation or acquisition of such metadata is crucial to their effective operation and ultimately user satisfaction. Knowledge representations, from formal ontologies to lightweight taxonomies and flat folksonomy-like term networks, are especially vital to advanced information processing tasks that need to process semantic relationships between entities. Typical examples of applications are metadata-based search engines or faceted browsers (Tvarožek & Bieliková, 2010), which require either document annotations or faceted classifications, (personalized) e-learning systems (Barla et al., 2010), which require complex course metadata, information repositories that need resource interlinks and annotations (e.g., Wikipedia), or Semantic Web applications that require formal ontologies. The use of concept relationships is also widely used in exploratory search tasks (Marchionini,

2006) like search query expansion (Ungrangsi, Anutariya, & Wuwongse, 2010) or visualization and navigation in information spaces (Stewart, Scott, & Zelevinsky, 2008).

Rather than use of semantics, more problematic is their acquisition. While automated approaches are able to provide quantity, in comparison with human oriented approaches (expert work, crowdsourcing) they vary in the quality of semantics they provide. The concept of *games with a purpose* (GWAP) emerged within the human computing initiative in recent years and stresses the use of human problem-solving capabilities via specially engineered games to address so called *human intelligence problems* (HITs) that are currently too difficult to solve by machine approaches (e.g., image annotation) (Siorpaes & Hepp, 2008a). By transforming computational problems into engaging computer games, GWAPs enable us to take advantage of human computing power without having to pay for expensive human resources while also providing the scalability to web-scale tasks. Several successful games like the image annotation ESP Game (Ahn & Dabbish, 2008) have already shown the potential of this approach, especially for creation of web semantics (Siorpaes & Hepp, 2008a).

However, the creation of GWAPs is not a straightforward process, because it is specific to each human intelligence problem that it tries to solve. Although some effort has been spent on developing generic methodologies for GWAP creation (Ahn & Dabbish, 2008; Siorpaes & Hepp, 2008a; Vickrey et al., 2008), they remain applicable only to narrow problem domains. In this paper, we examine design aspects of existing GWAPs (e.g., player input validation, anti-cheating, scoring system) – their shared characteristics and specifics to support the effort of devising a broader methodology, at least for the Semantic Web domain.

Our own contribution to the field of GWAPs for the Semantic Web includes *Little Search Game* – a novel game for the discovery of semantic links between terms and the successive construction of a term network. *Little Search Game* is a web search game, where players compete in reducing the number of search results by entering queries in a special format. The query consists of one normal search term, given to players before the game starts, and several negative terms entered by players, like "star –movie –wars –death". This query format forces players (in order to be successful in the game) to use negative terms with frequent common co-occurrences with the initial term (i.e., related to it). We collect this information via game logs and aggregate it into a lightweight term network of term relationships. The game has also the unique ability to discover term relationships as perceived by humans, some of which are hard to detect using statistical corpora analysis. Therefore the resulting lightweight term network is suitable for applications like learning support frameworks (Barla et al., 2010), exploratory search tools (Šimko, Tvarožek, & Bieliková, 2010) or as a subject for further semantic enrichment.

We present the details of the *Little Search Game*'s mechanics and term network construction. We further discuss the system of ladders and anti-cheating techniques employed. Next, based on our previous experiments, we briefly present quantitative results of the game deployment and validation of the acquired term network. Next, we discuss the attractiveness of the game with results of a player feedback survey. Lastly, we discuss the extension of the produced term network into a more ontology-like structure through naming of its relationships, and present experimental results examining the relationship types of the term network.

SEMANTIC ACQUISITION TECHNIQUES

Many approaches for knowledge and semantics acquisition have been developed. Automatic approaches to metadata creation address the requirements of web-scale information processing: they extract term relationships by mining text corpora using latent semantic analysis (Park &

Ramamohanarao, 2009) or by exploiting existing taxonomies (El Sayed, Hacid, & Zighed, 2007) or folksonomies (Barla & Bieliková, 2009), sometimes within specialized domains (Nenadić & Ananiadou, 2006; Šimko, 2011; Šimko & Bieliková, 2009). Products of such approaches (and also of our game) are term networks, similar to human-created folksonomies: graphs with terms or concepts as nodes, connected to others by relationships of a single type. More formally the network can be represented by a set of RDF triplets with only one possible predicate (which also raises the issue of posterior acquisition of true relationship types).

Other semantics acquisition approaches focus on the discovery of ontology triplets via harvesting of statements (sentences) within a text corpus. Approaches described by Pantel and Pennacchiotti (2008) or Sanchez (2010) extract triplets by searching for occurrences of predicates manually picked beforehand (general, e.g. "part of" or domain specific) and identifying subjects and objects they are bound with. An interesting "inversion" of this approach was developed by Weichselbraun et al., which labels already existing relationships. The method mines a text corpus looking for co-occurrences of terms coupled in unlabeled relationships and looks up for candidate predicates (Weichselbraun, Wohlgenannt, & Scharl, 2010).

While automatic methods are capable of delivering quantity (e.g., number of terms or concepts and relationships), they vary in quality (e.g., relevance of relationships) (Wang, Maynard, Peters, Bontcheva, & Cunningham, 2005) and often cannot cover nuances and unusual situations in corpora thus producing metadata that need additional validation by human experts. Human resources however are usually scarce and too expensive to be practically viable for web-scale tasks. Perhaps the most comprehensive manually created knowledge base and inference system today is CYC, which despite more than 30 years of concentrated human effort (more than a person-century) still has to achieve widespread practical acceptance and impact on the Web (Lenat, 1995). Consequently, the limited availability of metadata impedes the deployment of advanced applications and seriously hinders Semantic Web adoption.

GAMES WITH A PURPOSE

The basic premise of games with a purpose is that players willingly play a game and are rewarded by non-monetary values (e.g., fun), while they solve a corresponding problem as a side effect of the game (Krause, Takhtamysheva, Wittstock, & Malaka, 2010). To illustrate the potential of GWAPs for web-scale problem solving, we cite statistics provided by the Entertainment software association[1]. These indicate that about 273.5 million games were sold in 2009, corresponding to more than 550 million human-hours, if each game was played at least for two hours, still not considering games freely available on the Web (e.g., *Farmville* on Facebook). We believe that this sheer volume of untapped and potentially available human computing power could be partially harnessed to provide the resources necessary to create the semantic metadata required to bring the Semantic Web vision yet one step closer.

Despite the fact that games with a purpose present a recent development enabled by the Web and the number of online users as potential players, they have already been used to perform various tasks, mostly related to semantics and metadata acquisition. GWAPs take advantage of common features such as player consensus and their designers also have to address several common issues (game attractiveness, motivation, cheating prevention) in order to create a successful GWAP.

Aspects of GWAP design

The basic principle lies in *solving a computational problem via a carefully engineered game providing enjoyable experience to players*, who are usually unaware of its true purpose. Players engage in the game because it provides some non-monetary incentive, such as fun and entertainment. When playing, players' actions are logged and subsequently processed to generate useful artifacts e.g., metadata. The practicality of GWAPs for web-scale problems comes from their properties:

- massive parallelization due to the large number of potential players and their willingness to play,
- human intelligence can effectively solve many problems presently unsolvable by machine approaches,
- ability to achieve correct results even on noisy inputs via collaboration and agreement of multiple players.

Based on their properties, GWAPs present a unique opportunity to exploit human computation for the solving of complex, large scale problems virtually for free (relative to the size of the solved task). This is especially important in the Semantic Web scenario, which deals with huge amounts of information that needs to be processed by humans. From a design point of view, GWAPs share several issues and aspects that need to be considered during game design:

How player output is validated. The game concept has to overcome this paradox: the purpose of the game is to create new useful artifacts by tracking player actions, which only humans are able to do (otherwise the game is not needed), yet needs to validate the correctness of such artifacts automatically. Furthermore, this must be done immediately after the game ends to provide players with scoring feedback and motivate them to play more and create more useful artifacts. The winning conditions and useful artifact production must correlate, which GWAPs address in these ways:

- Ideally, a single-player approach implements techniques of automatic evaluation of the created artifacts (Terry et al., 2009). This is often strongly problem-specific and not applicable to most problems.
- Some games solve the issue using an online multiplayer scheme, where players validate each others' inputs (Ahn & Dabbish, 2008; Ho, Chang, Lee, Hsu, & Chen, 2009; Hladká et al., 2009; Siorpaes & Hepp, 2008a). However, this often results in the cold start problem due to the lack of players upon initial game deployment who would be willing to play at the same time. Many games solve this issue by introducing bot-players (Ahn & Dabbish, 2008; Siorpaes & Hepp, 2008a).
- In case of other single-player games, some GWAPs use bootstrapping techniques for result evaluation: they first test player reliability using problem (game) instances that have a known solution. They further mix the instances with known and unknown solutions to keep the player from creating invalid artifacts (Seneviratne & Izquierdo, 2010).

How entertainment is provided. According to Hunicke et al. players of computer games may be entertained by several game aesthetic factors (Hunicke, LeBlanc, & Zubek, 2004). In games with a purpose these usually include: (1) social experience – interaction with other players (Ahn & Dabbish, 2008; Tuulos, Scheible, & Nyholm, 2007), (2) self-challenge – overcoming a player's own previous achievement, joy of reaching the goal (Seneviratne & Izquierdo, 2010 ; Terry et al., 2009) and (3) competition among players – either in duels or via a ladder system (Hladká et al., 2009; Chamberlain & Poesio, 2009; Ho et al., 2009 ; Ahn & Dabbish, 2008).

How does the game prevent cheating attempts. Computer games, especially multiplayer and competitive, suffer from cheating attempts and dishonest player behavior (bypassing rules, exploiting bugs). In games with a purpose, cheating may not only destroy the fairness of the game but also cause invalid problem solutions to be generated. Solutions are often problem specific, i.e. the state space of the game can be analyzed and rule gaps identified. The common practice is also cross-validation among opponents (Ho et al., 2009). In the worst cases, beta testing of a game with live deployment quickly discloses problems with cheating.

The majority of the GWAPs rely on online multiplayer mode of gameplay to validate correctness of the players' actions and artifacts that are inferred based on their actions. The number of these games indicates the relative ease of implementing this principle, which also comes with the opportunity of implicit cheating prevention and potential social experience for players that seems to be a very strong incentive and motivation to play (Kuo et al., 2009).

GWAPs for Semantic Web

In the Semantic Web domain, GWAPs are being employed for various tasks ranging from resource annotation (multimedia or texts) to ontology building. A comprehensive summary of existing semantics acquisition GWAPs was created by Thaler et. al. (Thaler, Siorpaes, Simperl, & Hofer, 2011) and is maintained in the *Insemtives* web page[ii]. We distinguish two major groups of GWAPs based on their purpose: *resource annotation* – metadata acquisition games for multimedia (Ahn & Dabbish, 2008; Seneviratne & Izquierdo, 2010; Barrington, O'Malley, Turnbull, & Lanckriet, 2009; Ho et al., 2009; Dasdan et al., 2009), textual resources (Chamberlain & Poesio, 2009; Hladká et al., 2009) or web pages (Law, Mityagin, & Chickering, 2009), and domain modeling games, involving tasks like common sense facts collecting (Ahn & Dabbish, 2008; Kuo et al., 2009) and validation (Herdağdelen & Baroni, 2010), term associations acquisition (Vickrey et al., 2008) or ontology alignment (Siorpaes & Hepp, 2008a).

In the field of multimedia annotation, von Ahn's *ESP game* corresponds to an output-agreement game where two players must enter the same tag within a given time limit, given a common image as input and a set of taboo tags. Based on practical experience, as of 2008, 200,000 players have entered more than 50 million tags effectively solving the problem of image tagging via collaborative player agreement on image tags (Ahn & Dabbish, 2008).

Similarly to the *ESP game*, *KissKissBan* is an image labeling game that extends the original *ESP game* with additional anti-cheating aspects (Ho et al., 2009), since the original concept was prone to abuse if players agreed beforehand on the keywords they would input. While *ESP game* addressed this problem by selecting random players who ideally did not know each other or by using system level measures, *KissKissBan* extends the gameplay model with a "blocker" player whose task is to prevent the other two players from forming a unified strategy by monitoring the progress of the game and selecting taboo words on the fly.

Peekaboom is an inversion-problem game, which is designed to support the annotation of objects in images (Ahn & Dabbish, 2008). Again, two random players aim to find agreement – one plays the describer who is given an input, another one plays the guesser who must produce the input based on the description from the describer thus finishing the game.

As a non-multiplayer GWAP example working on the bootstrapping principle, the image annotation framework of Seneviratne and Izquierdo (2010) can be used. Here, a single player is annotating a pack of images with tags and receives his score only afterward. Some of the images in the set are already annotated and the scores is computed based on player's performance on those, plus the game uses some other heuristics for assumption whether the player is honest or

not. Players are aware of the scoring methods, so their best option to gain the highest possible score is to be honest all the time (Seneviratne & Izquierdo, 2010).

The bootstrapping approach for image annotation was also used in the *FishEye* game devised by Thaler et. al. although with a different technique of assigning annotations: instead of writing tags, players decide by putting images (fishes) into graphically stylized baskets, thus saying which of the given images belong to a specific concept assigned for each game round. This makes the task more comfortable for players (Thaler, Siorpaes, Mear, Simperl, & Goodman, 2011) and adds to effective "purpose encapsulation" exercised by the game's design (e.g. graphically stylized metaphor of fish catching).

Multimedia annotation GWAPs comprise mostly image annotation games. However, similar principles can also be used for annotation of video or audio streams. In the *Tagatune* game (Law, Ahn, Dannenberg, & Crawford, 2007), players need to agree whether they are listening to the same track or not by exchanging text messages (from which tag descriptions are extracted automatically). In *HeardIt* (Barrington et al., 2009), two players have to agree on the same values of certain attributes that describe an audio stream, which produces a (rather "supervised") track categorization. Similarly to that, Siorpaes and Hepp devised the *OntoTube* game for categorizing video streams (Siorpaes & Hepp, 2008b).

In case of textual resource annotation, the task of providing characteristic terms to a resource is successfully performed by automated natural language processing approaches. A more complex problem, suitable for human solving, is identification of co-references (matching nouns and pronouns in texts, referencing the same object). Two GWAPs were designed to fulfill this task: *PlayCoref* (Hladká et al., 2009) and *PhraseDetectives* (Chamberlain & Poesio, 2009). In *PlayCoref*, two players compete in marking co-references by matching nouns with pronouns. The score is afterward computed by validating player guesses against the opponent and also by comparison with the results of an automated co-reference detection approach (Hladká et al., 2009). *PhraseDetectives* is slightly different as it works in two rounds: annotation and validation (in which the opponents validate each others' guesses) (Chamberlain & Poesio, 2009).

In this paper, we describe a GWAP involving search query formulation by players. Some resource annotation GWAPs also utilize this feature. In *Thumbs-up,* the player, given two images and one query that retrieved them, has to decide which image suits the query better (Dasdan et al., 2009). In the game *Intentions*, two players have to agree, whether they see the same web page or not (each player retrieves one and cannot see the opponent's one). They reversely construct search queries to retrieve the original. They cannot see each other's queries, but may review other results from opponent's queries. The purpose here is to decorate initial web pages with relevant terms (Law et al., 2009).

Games with a purpose have also been employed in the field of domain modeling. Siorpaes and Hepp devised a set of interactive games for the Semantic Web whose purpose was the creation of annotations (*OntoTube*, *OntoBay*) but also ontology linking (*OntoPronto*) and ontology alignment (*SpotTheLink*) (Siorpaes & Hepp, 2008a; Thaler, Simperl, & Siorpaes, 2011). Several other GWAPs have been devised particularly in the field of ontology creation. *Verbosity* by Luis von Ahn, also an inversion-problem game, is played by two anonymous players where player B guesses a word (noun) given to player A. Player A describes the given word using other nouns inserted into preset sentence stubs either as subjects or objects. As players play, they discover common-sense facts about the words they play with – they connect them with relationships, typical for ontologies ("consists of", "is a", "is opposite of"), transformed to natural language sentence stubs. The correctness of the described relationships

stems from the fact that in the end, player B guesses the correct original word and also by cross-validation among game instances (Ahn & Dabbish, 2008). The game *Verbosity* is the main source of facts for the *ConceptNet* ontology.

Another ontology-population game is *OntoGalaxy* by Krause et al. (2010). In this single player game, designed graphically as an action game in space, players have to collect word-labeled freighters owing toward their ship, but only those labeled with a word satisfying certain conditions, given at the start of the game (e.g., "it must be touchable by hand"). The strengths of *OntoGalaxy* are its sophisticated graphics and good encapsulation of its purpose into its gameplay thus hiding it from players (Krause et al., 2010).

Concept game (Herdağdelen & Baroni, 2010) is an example of a GWAP oriented on the validation of common-sense facts obtained by automated means. Prior to gameplay, candidate triplet statements are acquired by an automated corpus-based method using seeds from the *ConceptNet* ontology. During the game, players encounter triplets being displayed by a slot machine (each feature on one of three cylinders). If a player thinks that a meaningful statement was rolled, he may "claim" money for it (effectively telling that he considers it valid). Based on other player choices, the player gets awarded or punished (Herdağdelen & Baroni, 2010). The game uses the bootstrapping approach of artifact validation, supported by heuristics indicating player's trustworthiness, similarly to Seneviratne's image annotation framework (Seneviratne & Izquierdo, 2010).

To summarize, games with a purpose have already shown significant potential to harness human computational capabilities for web-scale problems with respect to the cost of man-hours required. Although some authors formulated several best design practices (Siorpaes & Hepp, 2008a; Ahn & Dabbish, 2008), there still is no universal methodology for GWAP creation, which is often performed in an ad hoc way. Consequently, the transformation of arbitrary problems into a GWAPs, abuse detection and prevention, and on-demand metadata creation (e.g., how to focus GWAP to create/extend metadata in a desired way) are still challenging open problems.

LITTLE SEARCH GAME PRINCIPLES

In order to acquire a term relationship network (i.e., a light-weight semantic structure) we improve upon the state-of-the-art in games with a purpose by devising *Little Search Game*. It is a web search query formulation game, in which players reduce the number of results returned by a search engine to a minimum, using specially formatted queries (e.g., "star –movie –wars –death"), which force them to reveal their perception of term relationships. It is a single player, casual browser game that motivates players to play via ranked competition and mental challenges (game attractiveness factors defined by Hunicke et al. (2004)).

The game utilizes the principle of negative search, in which the original set of web search results is stripped of a subset of results containing specific negative terms, to construct a term relationship network by mining the game query logs. At the start of the game, the player is given a task in the form of a positive query term that yields a certain number of search results. The player's task is to reduce the number of results by adding proper negative terms to the given initial query term. The lower the final number of results, the better rank the player gets. In order to achieve the best results, players must enter negative terms that have high co-occurrences with the task term on the Web. This principle is also the key for term relationship networks acquisition since players interpret the co-occurrence of terms as a semantic relationship between them and vice versa (Šimko, Tvarožek, & Bieliková, 2011).

The winning condition of *Little Search Game* is evaluated automatically unlike in multiplayer-based games (Ahn & Dabbish, 2008; Ho et al., 2009; Law et al., 2007; Barrington et al., 2009) or bootstrap-based games (Seneviratne & Izquierdo, 2010; Thaler et al., 2011). Though the scoring only approximates the value of relationships created (best terms for the game are not necessarily best for the game's purpose and vice versa), it is sufficient to keep players motivated. This allowed us to design the game as a single player game. Consequently, the game does not suffer from the cold start problem caused by lack of players and neither requires a previously created set of data for score computation and result verification. Feedback to players is given immediately after the game ends. Since it cannot rely on mutual player control to prevent cheating, we use preventive rules to reduce dishonest player behavior and also impose a posterior cheating detection heuristics.

Game scenario

The following scenario illustrates a typical use of *Little Search Game* (the interface of the game is shown in Figure 1, the current implementation utilizes the Google search engine):

Figure 1. Example of the Little Search Game interface with negative terms (left), attempt history (center) and ranking ladder (right).

1. The player selects a task term or lets the game select one for him (we prefer to select tasks not yet played by the player), for example "star". The game queries a search engine with this term and displays the number of results, in this case nearly 500 million.
2. The player enters negative terms (fields on the left in Figure 1). He first uses the term "movie" (thus creating the query "star –movie"), because he feels that the words "star" and "movie" are commonly used within the same phrase in many web sites. After submitting this attempt, the search engine now returns only 400 million results. The change can be seen in Figure 1 in the middle of the game interface where the relative numbers of results per attempt (query) is shown.
3. The player may continue to refine the query by entering or overwriting the negative terms. He is limited to use up to N negative terms at once (we chose $N = 6$ to provide sufficient possibilities for players while not overloading them with too many options and to challenge the player to come up with the best terms). There is no penalty for performing more attempts; the players are free to refine their queries (and to enter new words) as much as they want. They usually do, until they are satisfied with their rank in the ranking ladder displayed on the right side of the interface (Figure 1).
4. When the player is satisfied with the results, he confirms the results and may review the rankings or play the same or another term in the next round. The ladder system is further discussed in section "Effectiveness of the game".

Term network inference

The format of the game queries forces players to reveal their perception of term relatedness with a given task term. If several players agree on the relatedness of the same term pair, we can arguably promote this relationship into the collaboratively created term relationship network (a subset of the *Little Search Game* term network is shown in Figure 2).

Figure 2. Subset of the created term network.

We represent log entries for term network creation as triplets (p_i, t_j, N_{ij}) where

- $p_i; p_i \in P$ is a player identifier where P is the set of all players.
- $t_j; t_j \in T$ is a task term where T is the set of all task terms.
- $N_{ij} = \{n_{ij1}, ..., n_{ijn}\}$ is a set of negative terms that player p_i used at least once when solving task t_j. Also let $\forall n; n \in N$ where N is the set of all negative words.

Note that there is no information about time, order, number of results for a particular attempt or the number of uses of a specific negative word. Though we also log this additional information, it is not needed for basic network creation. The only important fact is that the user used a negative term, which he at some point thought to be related to the task term (which it should occur often with). We call the first occurrence of the player, task term and negative term triplet a *vote* $l = (p_i, t_j, n_k)$ of vote set $L : P \times T \times N$.

If the number of votes with the same combination of task and negative term is greater than 5 (a constant we chose based on previous experience with such scenarios, i.e. we assume that the agreement of five players is satisfactory), the oriented relationship in the term network is created with the task word as the source node and the negative word as the target node. The relationship has two other attributes: the total number of votes that contain the source term as task word ω_t and the total number of votes for the particular relationship ω_p.

The *Little Search Game* term network is defined as a graph *G* consisting of the set of nodes *V* (containing task terms *T* and negative terms *N*) and set of edges *E*:

$$G(V, E)$$
$$V : T \cup N$$
$$E : V \times V \times \square \times \square$$

A term network edge $e; e \in E$ is a quartet $e = (t, n, \omega_t, \omega_p)$ where $t \in T$ and $n \in N$. Using values ω_t and ω_p we define the weight *w* of the edge relative to other edges outgoing from the same node:

$$w = \frac{\omega_p}{\omega_t}$$

The value *w* enables us to sort edges outgoing from a node by their relative strength.

Effectiveness of the game

To evaluate the problem solving potential of games with a purpose, Luis von Ahn suggested the use of *throughput* (the number of problem instances solved in one man-hour) multiplied by the *average lifetime play* (ALP) – the total number of hours dedicated to the game by one player. The ALP factor corresponds to the attractiveness of the game, which is crucial to maintaining player interest (Ahn & Dabbish, 2008). The game system influences both factors, which are usually contradictory.

In *Little Search Game* some of the attractiveness is motivated by the element of challenge (i.e., by the opportunity to outdo oneself), which is represented by a mental challenge for players

to come up with negative terms which really help them (Hunicke et al., 2004). The second part of attractiveness is competition (Hunicke et al., 2004). The challenge aspect is always present in the game, even if the player plays it alone. On the other hand, the competition depends on the overall fairness of the game (i.e., cheating detection and prevention handled primarily by word banning policy) and the ladder system.

The ladder is in fact the critical point in which the attractiveness and throughput contradict. From the attractiveness standpoint, all players should play the same set of task words, as that would be the only fair way to compare players (i.e., to compare their result counts for the same game tasks). But from the throughput standpoint, only a limited number of players is required for playing the same task, since we only need to collect enough votes to acquire the most relevant relationships in the term network (we opted for 10 relationships per term, this depends on the purpose of the network, for instance, for exploring the term network visually within exploratory search applications, 10 was sufficient). Allowing players to continue playing one task over and over again for the sake of competition would be inefficient, since new relationships would appear at a decreasing rate and be less relevant. Since we aimed to discover at least some relationships for many terms (and not many relationships for few terms), we devised a strategy for assigning task words effectively, without breaking the competition system.

As a compromise, we organize gameplay into rounds lasting from several days to weeks starting with an empty ladder. In each round, a set of task words is played and every task has a separate ladder, created as an ascending list of the achieved result counts (the best player has the lowest count). There is no penalty for playing the same task several times and all players can play all tasks, since we want to motivate players to play as much as possible to obtain more data. However, only the single best task result is recorded in the ladder.

We also devised an overall ladder (empty at round start), to which a player receives points depending on his ranks in individual task ladders. The first player in the task ladder receives 100 points to the overall ladder, the second 99 points etc. Consequently, players receive points even for low ranks (in terms of hundreds of players) and are motivated to play multiple tasks which significantly increases throughput.

Based on the traffic in the last game round, the length of the next round and number of tasks in the next task set is determined to optimize throughput and to generate the desired 10 relationships for each task term. However, the length of a round should not be shorter than about 3 days (to give more casual players enough time to compete) and the number of task words higher than about 20 (to make all tasks playable in a reasonable amount of time).

Purposeful task word selection

The strategy of selecting task words for LSG influences its outcome in several ways. Firstly, it enables us to control the "growth" of the term network: by introducing or ceasing the usage of certain terms, one can control the number of relationships created for that term. This can be exploited to complement similar existing term networks by "expanding" their nodes.

Secondly, task selection has impact on the attractiveness of the game. Players must be somewhat familiar with the task words they are given. If they do, they are going to be more effective in the game and therefore more satisfied. For general players, general terms should be used which yields general relationships. However, domain-specific scenarios are also possible. If the search query is made over a domain-specific corpus, indexed prior to gameplay (e.g., via tools such as *Lucene*[iii]) and the game is played by a group of users with some degree of expertise within this field, then it yields relationships specific for that domain. We observed this effect in

our experiments, where a significant number of players were students of IT: when they were given the task word “cellular” one of their negative term responses was “automaton” – a purely domain specific relationship.

A further extension of the game's concept can lie in the task selection strategy as the means of setting the difficulty of the tasks given to players to keep them challenged. The difficulty of the task correlates with the specificity of the task term (more specific = more difficult). Also, it can be modified by automated evaluation of past successes of other players with the task term (for example the relative decrease of the number of results).

Another possibility to acquire domain specific networks is to allow players to choose in which domain they want to play the game or assign tasks according to existing models of user interests (if available, for example in conjunction with a proxy server (Kramár, Barla, & Bieliková, 2010)). One of our aims is to use the game with groups of students during their courses (which act also as domains) and utilize the resulting relationships within the learning support system ALEF (Adaptive Learning Framework) (Barla et al., 2010) that uses lightweight semantics to model the syllabus of a course, helps students navigate through the course domain and is used for learning object recommendation.

GAME DEPLOYMENT

We devised *Little Search Game* as a web browser game (using the Silverlight platform) backed with .NET web services and the MS SQL database. Its current implementation is called *Little Google Game*[iv] as we used Google (via AJAX API) as the game’s web search engine. However, any search engine supporting negative term search is applicable. To access WordNet (for filtering only meaningful negative terms), we used the WordNet.Net library.

During the first deployment, the game was played by 30 players with an initial set of 20 arbitrary chosen task words (*Path, Bomb, Water, Castle, Cellular, Brain, Culture, Masquerade, Jaguar, Future, Star, Navy, President, Rontgen, Einstein, Easter, Worm, Beer, Forest, Sea*). Players played up to 300 games and submitted about 2000 queries. Even after such a small number of games, we were able to construct the base of a term network comprising more than 100 nodes of negative terms. To expand the term network, further task terms in later phases of deployment were picked from target terms of the strongest relationships in the graph.

Besides the casual game mode, we devised the tournament mode, which is suitable for organizing a game competition during various events (e.g., conferences). In the tournament mode, players solve different sets of tasks, but can also participate as regular players. We used the tournament mode during the Student Research Conference IIT.SRC 2010 at our university. Participants, mostly students, were motivated to participate via a tournament prize, but many of them played also the regular game tasks (off the prized competition).

Another release of the game was during a showcase, where a minor number of games was played. So far, the game was played by about 300 players with 3,800 played games and 27,200 submitted queries. The total number of task terms used in the game so far is 40 and players guessed over 3,200 negative terms.

The resulting network contains 400 nodes and 560 edges. However, the distribution of relationships per task term differs between game type in which it was used (tournament tasks were played much more than tasks used during the showcase). After imposing an additional log analysis restriction, that only the 10 strongest relationships per task term are considered, the resulting network shrunk to 183 nodes and 220 edges.

TERM NETWORK VALIDATION

We validated the acquired term network for semantic soundness. We conducted an experiment with a group of 18 judges (from various professions, aged 18-35) to evaluate the soundness of 20 term relationship (Šimko et al., 2011). Twelve of those relationships were from the network acquired using the game, 8 were created randomly (i.e. not sound). The relationships were shuffled so the judges were not able to figure out what they were expected to select. Judges were asked to assign each relationship a value between 1 ("definitely irrelevant") and 4 ("definitely relevant") to describe their opinion on the term pair's semantic soundness. They were also asked to evaluate pairs in an oriented manner, i.e. whether the term B is within the top ten most relevant terms for term A, but not necessarily vice versa.

After the judges submitted their votes, we computed their group opinion on soundness of each judged term pair. The outcome for each pair could be "sound", "not sound" and "controversial" in case that none of previous two options received more than two thirds of votes. To compute the number of votes, options 1 and 4 were counted with double weight (since the judges were more confident about them).

The experiment has shown that 10 out of 11 term pairs from the created term network were semantically sound, and one pair was not sound, which corresponds to almost 91% correctness. The twelfth relationship could not be evaluated, since it was marked as controversial – it was the pair "cellular – automaton" which makes sense for informatics (who were the prevailing group among players), but not in general.

HIDDEN RELATIONSHIPS

When considering methods for automatic term relationship extraction, term co-occurrence in documents comes to mind. However, the semantic soundness of term pairs does not necessarily correspond to the co-occurrence frequency of terms in large text corpora, such as the Web. For instance, the term pair "brain – tumor" is arguably a relevant connection, but those words, in fact, occur only sporadically together in web documents. On the other hand a pair "substance – argument" is a nonsense connection, despite that those terms occur ten times more frequently (according to their sole frequency of occurrence). This brings unwanted *noise* to co-occurrence based term network extraction, since the relevant pairs are "hidden" among irrelevant pairs with higher co-occurrence rates. The situation is best illustrated by the scheme on Figure 3.

Figure 3. Hidden term relationships – expectations (black) and reality (black and grey). Some semantically sound term relationships have their text corpus co-occurrence rates below the rates of nonsense term pairs and vice versa.

Little Search Game is able to identify these "hidden" term relationships. Although the players' winning goal is to identify term pairs with high co-occurrence rates, they interpret this as finding of terms relevant to the task term (which in their opinion also has the best chances for significant co-occurrence). Players thus sometimes enter negative terms that they *believe* will help them in reducing the number of search results but in fact do not significantly reduce the result set. If such terms are "hidden", they are quickly abandoned by players as useless, but remain in the game logs where they provide valuable information upon network extraction.

To validate, how many "hidden" term relationships our term network contains, we conducted an experiment, which examined the co-occurrence of term relationships present in the *Little*

Search Game (LSG) network (Šimko et al., 2011). Also, co-occurrences in a nonsense term pair set (created by random term pairing) were acquired to determine the "noise" level. We used the whole Web as a text corpus, accessed via the *Bing* search engine.

To determine the co-occurrence ratio of two terms, the search engine is queried for each term separately and then by conjunctive clause of both terms (e.g., "sea AND blue"). These three queries yield three result counts – p_s, p_t and i – which represent the cardinalities of result sets of the two terms and the intersection of those sets. The co-occurrence ratio of the search term to the target term is defined as $r_s = i / p_s$ (or $r_s = i / p_t$ for the reverse relationship).

We acquired the source-target co-occurrence ratios for the relationships in the LSG network and also for nonsense relationships in three reference sets. We used three sets composed of three different sized corpora of the most frequent words in the English language (800, 5,000 and 50,000 words), excluding stopwords. This was due to the fact, that more frequent terms, event semantically unrelated, produce higher level of noise. For our experiment, the relevant set was the medium sized (5,000 words) since it covered the terms used in the LSG network.

The measured values are plotted in Figures 4 and 5. If we look at the reference sets, we see that for the medium sized corpus, the noise starts to be significant at the ratio of 35%. Around 40% of the LSG term network relationships have co-occurrence rates below that value, which means they can be considered "hidden". The game is therefore able to discover a significant amount of relationships that cannot be retrieved by statistical corpora analysis.

Figure 4. Distribution of term network relationships by the real co-occurrence of the paired terms.

Figure 5. Distribution of the nonsense (reference) relationships by the real co-occurrence of the paired terms. Three differently sized corpora are used.

RELATIONSHIP TYPES

The LSG term network is a lightweight structure of untyped term associations. It is therefore relevant to discover, whether those relationships can be upgraded to more ontology-like triplets, for instance by assigning them appropriate types as predicates. Although the method of doing so is also the subject of our future work, we conducted experiments to examine the feasibility of such efforts. In the experiments, we again tested the soundness of the LSG term network relationships and examined which types of relationships are mostly present in the LSG network. We also examined the degree of possible enrichment of an existing knowledge base that we can achieve by adding relationships from our network. At the end of this section, we outline two possible scenarios how to perform relationship labeling over the LSG term network as our future work.

We have queried the existing general knowledge base of common facts – the *ConceptNet* ontology[v] – for types of LSG network relationships. We were interested in (1) how many of them are actually represented in such a knowledge base and (2) of what types they were. Additionally, we also conducted a manual (two judges 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 (defined by *ConceptNet*).

We picked the *ConceptNet* for several reasons. It contains common sense facts (incidentally, co-created by another GWAP, Verbosity (Ahn & Dabbish, 2008)) what makes it suitable to use

as a reference dataset. Second, it has a defined set of 23 possible general predicates, which allowed us to type the LSG network relationships with some precision, but has not dispersed them into too small categories.

Hypotheses were defined as follows: Relationships of the LSG are semantically sound and are of various types. In the existing knowledge base, most of the terms used in LSG network were also present in the form of concepts, but the knowledge base comprises only a limited number of (mostly rigid) LSG network relationships.

Data. 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. As knowledge base, we used *ConceptNet 3.0* accessible via an online REST API.

Process. 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 combinations. Judges also assigned one of the 23 relationship types (e.g., *IsA*, *HasA*, *UsedFor*, *CapableOf*) to each LSG relationship with an option to assign a default unknown type, which was also assumed if the judges had not reached an agreement. The automatic retrieval of relationship types from *ConceptNet* was straightforward.

Evaluation. We analyzed the collected data with these findings:

- The manual evaluation has shown that from 400 examined relationships 80% were semantically sound, 8% rather controversial and 12% not sound (which is less than in the previous experiment, however, the "strongest" 100 relationships still had 93% soundness with only one 1% marked as not sound).
- The *ConceptNet* comprises only 164 (41%) out of 400 relationships we worked with. We consider this to be a strong argument for putting effort into enriching such knowledge bases with relationships acquired by *Little Search Game*.
- The distribution of relationship types in the manually annotated set and *ConceptNet* set differs (as shown in Figure 6). First, many of the *ConceptNet* evaluated relationships were taxonomic (IsA) while various dependencies (Desires, Causes...) were virtually absent in *ConceptNet*. Generally, *ConceptNet* relies roughly on 6 types of relationships, while the manually annotated set appears to be much richer in types what further stresses the usefulness of naming LSG created relationships.

Figure 6. Relative distribution of relationship types in the manually evaluated set and ConceptNet.

Verb retrieval via Little Search Game

The basic game principle of *Little Search Game* can be modified to serve the relationship labeling purpose: (1) The task query can comprise two nouns (of the unnamed relationship) instead of single one. (2) The possible negative terms could be restricted to verbs.

The rest of the game rules remain untouched. The advantage of this modification is that players do not have to learn a new game principle; they just switch the game to different mode. Since the players (as demonstrated earlier) think about game tasks on a conceptual level, we argue that they will (as a first choice) use verbs effectively describing some of the valid predicates between the given terms. A minor drawback of this approach is the bypassing of the sentence syntax the player has in mind and the introduction of ambiguity in deciding what is the left or right feature in the triplet, however, even the assignment of a predicate is valuable.

Mining sentences on the Web

The second method that we propose for labeling the relationships of the LSG network relies on mining text for triplet-like statements comprising terms of the unnamed relationship. The main idea is to retrieve relevant web (or domain corpora) documents for each unnamed relationship using a web search engine, then parsing sentences containing relationship terms and finally analyzing those sentences for triplet statements. A similar approach was already implemented by McDowell and Carafella for unsupervised instance population for ontology classes and exploring relationships (Mcdowell & Cafarella, 2008).

In this method, the system of result ranking implemented by search engines is exploited. Since it takes into account the co-location of terms within the documents, it can retrieve relevant web documents to mine for relatively small cost: the most relevant documents will rank high in the result list. We propose sentence mining in several ways:

- Three-term sentences. A huge corpus of potentially relevant documents allows us to look only for three-term sentences, containing always two of our relationship terms and a third term which will be considered as a third part of a triplet. With the use of part-of-speech identification (e.g., using WordNet) we could identify predicates for noun-noun bigrams (common case of unnamed relationships), but also nouns complementing the ”noun-predicate“ which are sometimes also present in the LSG network.
- Using the option above, we could search for any label (type) for a relationship. We could however, focus only on a limited set of general types of relationships as they are in many cases sufficient (e.g., *IsA*, *HasA*, *LocatedNear* as seen in *ConceptNet*). The lookup can then be performed using the predefined set of manually created sentence stubs characteristic for expressing the relationship (e.g., “A such as B” for “IsA” relationships).

BANNED TERMS

Game rules must be defined to preserve fairness and eliminate opportunity for abuse (besides “hard line” hacks, such as attacking network communication or the software itself). Abuse is usually done by a small group of (somehow curious) players, but especially in case of games with ladders, these players can ruin the whole system of fairness and cause regular players to abandon a game. With respect to games with a purpose, rules also have to eliminate possibilities for player behavior that does not support the game's purpose (e.g., not solving the problem correctly).

In *Little Search Game* the only player action that affects the game outcome are the negative term entries. After a few games, curious players realized that entering stopwords either from a language standpoint (e.g., “star –is –the –over”) or Web standpoint (e.g., “star –download –page –table –menu”) yields very few search results for all search terms and thus promotes them to top ranks in ladders with scores impossible to achieve by “proper” means. Some of the “web stopwords” like “menu” or “download” originate from their widespread use on websites, others like “page” or “table” occur widely within HTML source code. As search engines include them into indexes, they bring an unwanted bias into the game.

We addressed this issue of “killer words” by defining a set of terms allowed to be used in the game, before the game was released:

- The terms must be contained within a dictionary (we used *WordNet*[vi]) within one of its parts of speech: nouns, adjectives, verbs or adverbs (i.e., we exclude most semantic-less terms). Before a game attempt is sent to the search engine, the negative terms are checked against the dictionary, which also helps us detect misspelled terms.
- Using the same term as the task term is forbidden; we also use *WordNet* to check for morphs of the task term.
- Most frequent words in the English language (the language of the game) are excluded. We used the set of the 200 most frequent words in the texts of Wikipedia.
- HTML tags and some other words common on the Web are excluded. This manually prepared set of terms was extended with additional words that were banned from use during the first days of game deployment.

Though we proactively devised these rules, some of the non-suitable words were only discovered after having been abusively used by players after the game was released. Thus we devised a term abuse detection heuristic that suggests suspicious words to game administrators (who may eventually decide to ban additional terms). The premise is that possibly abusive words would appear in the attempts of leading players and eventually be used multiple times for different tasks and their result yield would be significantly lower than the nearest ranking attempts.

Our heuristic works in these steps:

1. Collect (three) highest ranking players for each task.
2. Exclude players that rank in just one task ladder.
3. Collect negative terms for each of the remaining players' best attempts in tasks where they ranked high.
4. Exclude terms appearing in only one task.
5. Mark terms that appear in more than two tasks or that are contained in attempts yielding significantly fewer results (more than 10% than the next ranking attempt) as suspicious and pass them to an administrator.

We compute the "universal effectiveness" for suspicious terms as a guideline to administrators (to determine if they are truly stopwords) by querying for their co-occurrence within a set of 10 manually selected reference terms (with a small mutual co-occurrence on Web). If the intersection with them is high, the terms should probably be banned.

Formally: let $S = \{s_1, ..., s_n\}$ be a set of suspicious terms. For each s_i we construct a set $P_i = \{p_{i1}, ..., p_{im}\}$ of ordered pairs $p_{ij} = (r_j, q_{ji})$ where $r_j \in R$ is a reference term belonging to the set of reference terms $R = \{r_1, ..., r_m\}$ and q_{ji} is a search query of reference term r_j and the suspicious term s_i formatted as "r_j AND s_i". The search engine when queried by elements of each ordered pair p_{ij} yields elements of another ordered pair (σ_j, ψ_{ij}) where $\sigma_j; \sigma_j \in$ □ is the number of results yielded for the reference term r_j only and $\psi_{ij}; \psi_{ij} \in$ □ is the number of results yielded for query q_{ji}. Relatively high values of ψ_{ij} against σ_j (more than 30%) in the majority of reference terms, argued strongly for the banning of terms, though the final decision was made by game administrators.

After the game was released, the list of banned terms has increased from around 230 terms to 430. Using our heuristic we discovered about 30 non-suitable terms; the rest was manually inferred. Sometimes, one abused term indicated potential use of other similar terms. One example of discovered abusable terms were numbers and digits as some are present in WordNet.

Since players received no penalty for trying to find "killer words", we have observed repeated attempts to do so. In fact, we welcomed their efforts as they helped us make the game fairer.

Attempts rejected due to use of banned words were returned to players with an explaining message – players were aware of all the aforementioned criteria. Unfortunately, some terms had to be excluded from the game (and thus the resulting term network) even though they arguably had semantic meaning and legitimacy to be used as negative terms in certain tasks (e.g., "restaurant –menu" or "school –table", where "menu" and "table" are banned due to being common in web content and HTML code). Such terms had to be sacrificed in order to keep the game fair; though they could still be used as task words.

GAME ATTRACTIVENESS AND EXPLICIT USER FEEDBACK

Game attractiveness plays an important role in evaluating the game's problem solving potential. Luis von Ahn mentions the average lifetime play factor (Ahn & Dabbish, 2008), which depends mostly on a game's ability to maintain player interest in the game by providing fun. However, the overall game potential is also dependent on the game's ability to spread and attract new players. Thus to evaluate our game from the attractiveness point of view, we conducted the following player survey.

Environment. The survey was conducted during a showcase venue called Researchers' Night, which took place in a shopping mall where visitors had the opportunity to observe or try out various scientific experiments in a joyful and popular way. This included the *Little Search Game*, located in one of the showcase kiosks. The game principles as well as its purpose were explained with the help of a poster.

Participants. Visitors of the showcase were aged 14 years or older but comprised mostly adults. From about 70 players, 34 voluntarily completed the questionnaire after they had played several games. All of the participants had previous experience with web search, but approximately half of them were not aware of the negative search feature.

Data. We integrated a built-in questionnaire into the *Little Search Game* interface to pair player answers with their gaming sessions.

Task. Players were asked to choose one of the given answers to questions shown in Table 1. All participants answered all questions. The questions focused on whether and how players understood the game and its purpose, their general attitude to games and particularly the *Little Search Game*, the game's potential for viral spreading and their attitude towards the game's purpose. Participants could also fill in an optional free-text field to explain their answers. We also collected some feedback during informal interviews with the participants.

The result summary of the survey is shown in Table 1. From the attention attraction standpoint, the game was not self-explanatory: less than half of the players understood it only after reading the manual. Informally, some players mentioned that even with the manual, they would not be interested in the game if they came across it, because they prefer a self explaining interface. Not surprisingly, the non-awareness of the concept of negative search was also a drawback for players. However, once it was clear, players were more willing to play the game repeatedly.

Table 1. Results of the attractiveness survey from the Little Search Game showcase experiment.

Question/Answer	**Answer count**
1. Have you understood the game?	

No.	2
Yes, after reading the manual.	16
Yes, after spoken explanation by the game's author.	16
2. Have you understood the purpose of this game (acquisition of term relationships)?	
No.	4
Yes.	30
3. Do you play computer games?	
Not at all.	7
Casually.	18
Regularly.	9
4. Would you play Little Search Game again?	
One time was enough.	6
Yes, I would like to play again.	18
Yes, I would like to play again and would recommend it to my friends	10
5. If you did not know that the game had an useful purpose, would you play it?	
I would play it anyway.	29
If the game was not useful, I would not play it.	5

On the other hand, once players understood the game principles, they considered the game interesting (question number 4 was indirectly aimed at the attention-keeping aspect of the game). Around 82% of participants expressed their interest and nearly one third of them would recommend the game further, which shows the game's viral spreading potential.

In the survey, we also asked about the perception of the game's purpose. First, we asked whether the players understood the purpose at all, which came out in correlation to understanding the game (since the explanation of the purpose was an integral part of the explanation of the game). Then we asked whether the purpose was the only motivation to play the game. A small, though significant number of players responded yes, which means that altruism may help overcome the initial barrier of lacking game attractiveness and argues for disclosing the purpose of the game to potential players. Note, that this does not mean that the players must feel the purpose in game (through rule deformations), which may lead to disturbance of players as shown by Krause et al. (2010).

CONCLUSIONS AND FUTURE WORK

The creation and maintenance of web-scale semantic metadata (e.g., ontologies or lightweight semantic networks) has so far been an open problem. Games with a purpose (GWAP), a paradigm to human computing pioneered by Luis von Ahn (Ahn & Dabbish, 2008), emerged as an alternative way to provide solutions to complex computational problems. They employ the power of human minds, superior to machine computation in many tasks (such as knowledge delivery, extraction and reasoning), while retaining the quantitative capabilities of web-scale computation through mass parallelization.

We devised and evaluated a novel approach to term network acquisition – *Little Search Game*, which effectively combines a GWAP with a web search engine to address collaborative semantic metadata acquisition. We described the basic premise behind *Little Search Game* in

which players extend the initial query with negative terms in order to reduce the total number of results as much as possible thus supplying us with their perception of term relatedness. The game is single player and thus does not suffer from the cold start problem. Our evaluation has conclusively shown that:

- *Little Search Game* can be used to create very precise term networks with relatively little human effort.
- *Little Search Game* can be used to identify term relationships which are normally not discoverable by statistical analysis of web corpora due to low co-occurrence of the respective terms. This can be achieved by exploiting player perception of term relatedness instead of statistical text analysis which cannot identify these relationships due to noise.
- The game-created term relationships are of various types and are good candidates for extension into ontology triplets and incorporation into existing knowledge bases.
- The proposed heuristic can be effectively used to identify potentially abusive terms used by players thus lowering administrative workload of game administrators.
- 82% of players were interested in the game and played it despite some drawbacks in the intuitiveness of its user interface. Almost 30% of players were inclined to recommend the game to others supporting its viral spreading to new players.

Future work will focus on the enhancement of several game aspects. We plan to increase the game's attractiveness by improving its interface to be more intuitive and by employing it in a more social way to engage players in the competition. We plan to integrate a messaging mechanism that will notify players about ranking changes and motivate them to respond to them.

Labeling of the relationships of the LSG term network is also one of our future goals. Our aim is to employ textual corpora mining techniques seeking for sentences containing bigrams of the term network and extracting third elements of the contained statements to form triplets. We also plan to introduce a modification of the LSG, where two nouns (already connected within LSG network) act as a task query. Allowed only to enter verbs as negative search terms, players would effectively label unnamed relationships with predicates.

We also aim to develop strategies for term network expansion, where the goal is to cover as many terms as possible with focused expansion of the most “relevant” terms for a particular task, where the term network is employed. We plan to create relationships “on demand”, also considering relationship strengths that can be derived from additional logged data (e.g., order of negative terms). For example, we want to support search query expansion of the most common queries and use the game in conjunction with our adaptive proxy-server for web search and browsing, where we collect bags of words via collecting and mining user search logs (Kramár et al., 2010). Here, term relationships help with grouping of user profiles or content-based recommendations.

ACKNOWLEDGEMENT

This work was supported by grants No. VG1/0675/11, APVV 0208-10 and it is a partial result of the Research and Development Operational Program for the projects Support of Center of Excellence for Smart Technologies, Systems and Services, ITMS 26240120005 and Research and Development Operational Program for the projects Support of Center of Excellence for Smart Technologies, Systems and Services II, ITMS 26240120029, co-funded by ERDF.

The authors wish to thank all members of the Personalized Web research group (pewe.fiit.stuba.sk) for their help with experiments.

REFERENCES

Ahn, L. von, & Dabbish, L. (2008). Designing games with a purpose. *Communications of the ACM, 51*, 58-67.

Barla, M., & Bieliková, M. (2009). On deriving tagsonomies: keyword relations coming from crowd. In N. T. Nguyen, R. Kowalczyk, & S.-M. Chen, (Eds.) *Lecture Notes in Computer Science* (vol. 5796, pp. 309-320). Springer Berlin Heidelberg.

Barla, M., Bieliková, M., Ezzeddinne, A. B., Kramár, T., Šimko, M., & Vozár, O. (2010). On the impact of adaptive test question selection for learning efficiency. *Computers and Education, 55*(2), 846-857.

Barrington, L., O'Malley, D., Turnbull, D., & Lanckriet, G. (2009). User-centered design of a social game to tag music. In *Proceedings of the ACM SIGKDD Workshop on Human Computation* (pp. 7-10). New York, NY, USA: ACM.

Chamberlain, J., Poesio, M., & Kruschwitz, U. (2009). A demonstration of human computation using the phrase detectives annotation game. In *Proceedings of the ACM SIGKDD Workshop on Human Computation* (pp. 23-24). New York, NY, USA: ACM.

Dasdan, A., Chris, D., Kolay, S., Alpern, M., Han, A., Chi, T., . . . Verma, S. (2009). Thumbs-up: a game for playing to rank search results. In *Proceedings of the ACM SIGKDD Workshop on Human Computation* (pp. 36-37). New York, NY, USA: ACM.

El Sayed, A., Hacid, H., & Zighed, D. (2007). Mining semantic distance between corpus terms. In *Proceedings of the ACM first PhD Workshop in CIKM* (pp. 49-54). New York, NY, USA: ACM.

Herdağdelen, A., & Baroni, M. (2010). The concept game: Better commonsense knowledge extraction by combining text mining and a game with a purpose. In *Proceedings AAAI Fall Symposium on Commonsense Knowledge* (pp. 52-57). Palo Alto, California: AAAI Press.

Hladká, B., Mírovský, J., & Schlesinger, P. (2009). Designing a language game for collecting coreference annotation. In *Proceedings of the 3rd Linguistic Annotation Workshop* (pp. 52-55). Stroudsburg, PA, USA: Association for Computational Linguistics.

Ho, C.-J., Chang, T.-H., Lee, J.-C., Hsu, J. Y.-j., & Chen, K.-T. (2009). KissKissBan: a competitive human computation game for image annotation. In *Proceedings of the ACM SIGKDD Workshop on Human Computation* (pp. 11-14). New York, NY, USA: ACM.

Hunicke, R., Leblanc, M., & Zubek, R. (2004). MDA: A formal approach to game design and game research. In *Proceedings of the Challenges in Games AI Workshop, Nineteenth National Conference of Artificial Intelligence* (pp. 1-5). Palo Alto, California: AAAI Press.

Kramár, T., Barla, M., & Bieliková, M. (2010). Disambiguating search by leveraging the social network context based on the stream of user's activity. In *Proceedings of the 18th International Conference on User Modeling, Adaptation, and Personalization* (pp. 387-392). Hawaii, HI, USA: Springer.

Krause, M., Takhtamysheva, A., Wittstock, M., & Malaka, R. (2010). Frontiers of a paradigm: exploring human computation with digital games. In *Proceedings of the ACM SIGKDD Workshop on Human Computation* (pp. 22-25). New York, NY, USA: ACM.

Kuo, Y.-l., Lee, J.-C., Chiang, K.-y., Wang, R., Shen, E., Chan, C.-w., & Hsu, J. Y.-j. (2009). Community-based game design: experiments on social games for commonsense data collection. In *Proceedings of the ACM SIGKDD Workshop on Human Computation* (pp. 15-22). New York, NY, USA: ACM.

Law, E., Mityagin, A., & Chickering, M. (2009). Intentions: a game for classifying search query intent. In *Proceedings of the 27th International Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 3805-3810). New York, NY, USA: ACM.

Law, E. L., Ahn, L. von, Dannenberg, R. B., & Crawford, M. (2007). Tagatune: A game for music and sound annotation. In *Proceedings of the 8th International Conference on Music Information Retrieval* (pp. 361-364). Vienna, Austria: Austrian Computer Society.

Lenat, D. B. (1995). Cyc: a large-scale investment in knowledge infrastructure. *Communications of the ACM, 38*(11), 33-38.

Marchionini, G. (2006). From finding to understanding. *Communications of the ACM, 49*(4), 41-46.

McDowell, L. K., & Cafarella, M. (2008). Ontology-driven, unsupervised instance population. *Web Semantics, 6*, 218-236.

Nenadić, G., & Ananiadou, S. (2006). Mining semantically related terms from biomedical literature. *ACM Transactions on Asian Language Information Processing (TALIP), 5*(1), 22-43.

Pantel, P., & Pennacchiotti, M. (2008). Automatically harvesting and ontologizing semantic relations. In *Proceeding of the Conference on Ontology Learning and Population: Bridging the Gap Between Text and Knowledge* (pp. 171-195). Amsterdam, The Netherlands: IOS Press.

Park, L. a. F., & Ramamohanarao, K. (2009). An analysis of latent semantic term self-correlation. *ACM Transactions on Information Systems, 27*(2), 1-35.

Sanchez, D. (2010). A methodology to learn ontological attributes from the web. *Data & Knowledge Engineering, 69*(6), 573-597.

Seneviratne, L., & Izquierdo, E. (2010). An interactive framework for image annotation through gaming. In *Proceedings of the International Conference on Multimedia Information Retrieval* (pp. 517-526). New York, NY, USA: ACM.

Siorpaes, K., & Hepp, M. (2008a). Ontogame: weaving the semantic web by online games. In *Proceedings of the 5th European Semantic Web Conference on the Semantic Web: Research and Applications* (pp. 751-766). Berlin, Heidelberg: Springer-Verlag.

Siorpaes, K., & Hepp, M. (2008b). Games with a Purpose for the Semantic Web. *IEEE Intelligent Systems, 23*(3), 50-60.

Stewart, R., Scott, G., & Zelevinsky, V. (2008). Idea navigation: structured browsing for unstructured text. In *Proceedings of the 26th SIGCHI Conference on Human Factors in Computing Systems* (pp. 1789-1792). New York, NY, USA: ACM.

Šimko, J., Tvarožek, M., & Bieliková, M. (2010). Semantic history map: Graphs aiding web revisitation support. (2010). In *Database and Expert Systems Applications, International Workshop on Web Semantics* (pp. 206-210). Los Alamitos, CA, USA: IEEE Computer Society.

Šimko, J., Tvarožek, M., & Bieliková, M. (2011). Little search game: term network acquisition via a human computation game. In *Proceedings of the 22nd ACM conference on hypertext and hypermedia* (pp. 57-62). New York, NY, USA: ACM.

Šimko, M. (2011). Automated domain model creation for adaptive social educational environments. *Information Sciences and Technologies Bulletin of the ACM Slovakia, 3*(2), 119-121.

Šimko, M., Bieliková, M. (2009). Automatic concept relationships discovery for an adaptive e-course. In T. Barnes, M. Desmarais, C. Romero, S. Ventura (Eds.) *Proceedings of Educational Data Mining 2009, 2nd International Conference on Educational Data Mining* (pp. 171-179). Cordoba, Spain: EDM.

Terry, L., Roitch, V., Tufail, S., Singh, K., Taraq, O., Luk, W., & Jamieson, P. (2009). Harnessing human computation cycles for the FPGA placement problem. In *Proceedings of ERSA* (pp. 188-194). Las Vegas, Nevada, USA: CSREA Press.

Thaler, S., Siorpaes, K., Mear, D., Simperl, E., & Goodman, C. (2011). SeaFish: A game for collaborative and visual image annotation and interlinking. In G. Antoniou, M. Grobelnik, E. Simperl, B. Parsia, D. Plexousakis, P. De Leenheer, & J. Pan (Eds.), *The Semantic Web: Research and Applications* (Vol. 6644, p. 466-470). Berlin, Heidelberg: Springer Verlag.

Thaler, S., Siorpaes, K., Simperl, E., & Hofer, C. (2011). *A survey on games for knowledge acquisition* (Tech. Rep. 2011-05-01). Innsbruck, Austria: University of Innsbruck.

Tvarožek, M., & Bieliková, M. (2010). Generating exploratory search interfaces for the semantic web. In P. Forbrig, F. Paternó, & A. Mark Pejtersen (Eds.), *Human-computer interaction* (Vol. 332, pp. 175-186). Boston, USA: Springer.

Ungrangsi, R., Anutariya, C., & Wuwongse, V. (2010). Enhancing folksonomy-based content retrieval with semantic web technology. *International Journal on Semantic Web and Information Systems, 6*(1), 19-38.

Vickrey, D., Bronzan, A., Choi, W., Kumar, A., Turner-Maier, J., Wang, A., & Koller, D. (2008). Online word games for semantic data collection. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 533-542). Morristown, NJ, USA: Association for Computational Linguistics.

Wang, T., Maynard, D., Peters, W., Bontcheva, K., & Cunningham, H. (2005). Extracting a domain ontology from linguistic resource based on relatedness measurements. In *Proceedings of the 2005 IEEE/WIC/ACM international conference on web intelligence* (pp. 345-351). Washington, DC, USA: IEEE Computer Society.

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

---

[i] http://www.theesa.com/facts/index.asp
[ii] http://www.insemtives.eu/games.php
[iii] http://lucene.apache.org/
[iv] http://mirai.fiit.stuba.sk/LittleGoogleGame
[v] http://csc.media.mit.edu/conceptnet
[vi] http://wordnet.princeton.edu