WikiWarp: Human Computation for Semantic Relevance

Nikhil Srivastava

CS 286r Final Report

Abstract

The semantic relationship between arbitrary human concepts contains information invaluable for information retrieval and natural language processing but notoriously difficult to teach computers. The situation is suitable for a *game with a purpose* whose human participation results in the solving of computational tasks that are difficult for computers alone to perform. Here I present WikiWarp, a computation system in the form of an online GWAP designed to extract semantic relatedness between concepts by aggregating and processing user-navigated paths through Wikipedia's link graph. The game's motivation, design and implementation are described, and a preliminary evaluation is performed of its enjoyability and usefulness based on data collected in a pilot study.

1 Introduction

A *game with a purpose* (GWAP) is a game played by human participants whose gameplay aids in performing computational tasks that are difficult to automate or perform efficiently with computers alone. Traditional computational targets include image [1] or audio [4] labeling and aggregation of common sense knowledge [5]. Instances of these games have proven to be quite popular: Verbosity, Tag-a-Tune, and the ESP game currently collect data (at www.gwap.com), and the latter has been adapted for Google's Image Labeler.

In this paper a novel GWAP is presented, one focused on the computational task of determining the semantic relatedness of different concepts. In the hopes of acquiring data with this semantic content, the game encourages users to navigate between prescribed start and target pages using only Wikipedia's internal page links. The collected data is thus an aggregation of directed walks through the Wikipedia link graph.

This paper first provides a motivation for the WikiWarp game by discussing the problem of semantic relatedness that is its computational target. Next, it presents the game's design and implementation as the result of a series of iterative improvements. Finally, it evaluates the game in terms of its enjoyability and usefulness, making a preliminary attempt to validate a subset of the data generated in the pilot study.

2 Motivation

A class of computational problems that has been traditionally very hard for computers to solve is measuring the semantic relatedness between human concepts. Humans are naturally adept at communicating and processing the relationship in meaning between two arbitrary concepts, such as "Bathtub" and "Hygiene", but representing this information and effectively teaching it to computers remains a major obstacle. Overcoming such a challenge is thought to be invaluable in the development of next-generation information retrieval algorithms and general progress in the fields of natural language processing and artificial intelligence.

In part, the insurmountability of the task arises because the problem involves two steps - acquiring large sets of data and extracting their semantic content - that are notoriously difficult to perform at the same time. Some strategies, such as statistical

analyses of large textual data sets, generate a large volume of information but with questionable (or difficult to extract) semantic content; other methods, such as those that directly enlist humans, provide rich semantic content but require the completion of tedious tasks that cannot generate large volumes of data quickly and cheaply.

Thus there is a viable opportunity to exploit human computing in the form of a GWAP to acquire this set of information. One possible implementation is a game in which players are placed on a *start page* on Wikipedia and instructed to navigate through the internal links of the website to arrive at a specified *target page*. The paths constructed between disparate topics contain some information about their semantic relatedness, and the goal of the game is to encourage the high-volume creation of this user-generated information. The hypothesis is that effective processing of this information might provide a layer of semantic relatedness unobtainable by previous measures, such as via link distances or linguistic similarity.

3 Motivated Design

Unfortunately, most of the GWAPs in existence are multiplayer or player-against-bot games that share certain input/output templates, which makes much of the prior detailed analysis of these games unsuitable or irrelevant to the design of WikiWarp. (Though below, I discuss the possibility of adapting WikiWarp to be more like traditional GWAPs). Nevertheless, many of the considerations relevant to promoting enjoyability, increasing output, and preventing cheating are applicable.

Most important, however, is perhaps the cardinal rule in game development: *gameplay should be structured so that it is each player's optimal strategy to successfully aid the computation that the game is designed to complete.* Much of the subtlety and finesse in game design is aligning the best interests of participant with those of the game designer:

Verbosity

Optimal Player Strategy	guess/reveal word as quickly as possible
Computational Goal	identify common-sense facts
Modifications	prohibit rhyming words (avoid "sounding-out" strategy) prohibit short words (avoid "spelling-out" strategy) randomly match players (avoid meta-game strategies)

ESP Game

Optimal Player Strategy	match keywords as quickly as possible
Computational Goal	generate accurate keywords
Modifications	prohibit taboo words (already known keywords) prohibit colors (generic, unhelpful keywords) randomly match players (avoid meta-game strategies)

WikiWarp

Optimal Player Strategy	navigate to target as quickly as possible
Computational Goal	generate "meaningful" path between two concepts
Modifications	prohibit irrelevant lists - e.g. "List of deaths in 2004" prohibit computationally shortest path(?) ?

Figure 1: Optimal participant strategy versus computational goal for some popular GWAPs.

For WikiWarp, an additional complication is posed by the fact that optimal gameplay *does not* solve a problem that a computer cannot. In fact, the situation is quite the opposite - shortest-path algorithms for graph traversal can identify the quickest route between prescribed pages with little difficulty. The problem is that these solutions have little content relevant to the semantic relatedness between the topics: the shortest path between "Mount Everest" and "Camel" is simply "Mount Everest" - "Geography of

China" - "Camel". Often, arbitrary lists such as lists of accidents or birthdays or days of the year are used to connect disparate topics. Worse, algorithms exist that make these routes freely available online [1]. Thus "optimal" gameplay is decidedly not in the designer's best interests. Such strategies can be dissuaded (by eliminating generic list pages) or made illegal (by identifying and prohibiting the freely available "optimal" route), but such restrictions will quite possibly infringe on the natural navigation of honest users.

4 Game Design - Gameplay and Implementation

The final implementation enabled users to choose between three game types:

- **featured game** with start and target randomly chosen from a database of 30 pre-selected games,
- **custom game** with user-selected start and target pages, and
- **random game** with start and target pages randomly chosen using Wikipedia's random article function

where the latter two were implemented largely due to user suggestions. To prohibit users from repeatedly playing excessively easy custom games, this game type was unscored. Random games, due to their higher difficulty, did reward users with higher scores for the same time interval and link distance; however, the increase was evidently not enough to offset the difficulty because random games accounted for less than 3% of total. As expected (and desired), the featured games accounted for 94% of all completed games and will be the focus of this paper.

Once start and target pages have been determined, gameplay is incredibly simple. At each stage of each game, the player is presented with a yellow status bar listing their current page, their target page, the time elapsed and links traversed, current score (the score they would achieve if they ended the game at that moment), and a parsed representation of the Wikipedia page they are currently on. Parsing simply selects the body of the article, removes navigation boxes, extraneous image tags and "list of" pages, and associates each internal link with a unique hash that prevents users from simply typing in the target URL to complete the game. A player is allowed to click on any of the links they believe will move them closer to their target.

When a player navigates successfully to the target page, they are presented with their path and final score, their comparative performance against other completions of the same game, and their cumulative performance if they have registered on the website. The score is currently computed solely based on time - monotonically decreasing from 1000 points at start to roughly 200 points at 500 seconds, but scores based on path length are easily implemented. Screenshots from a game in progress and a completed game are shown in Figures 2 and 3.

[1] Six Degrees of Wikipedia - http://www.netsoc.tcd.ie/ mu/wiki/

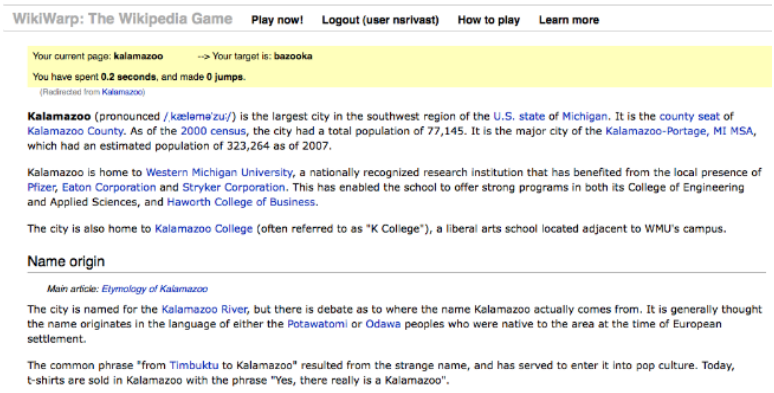

Figure 2: A game in progress.

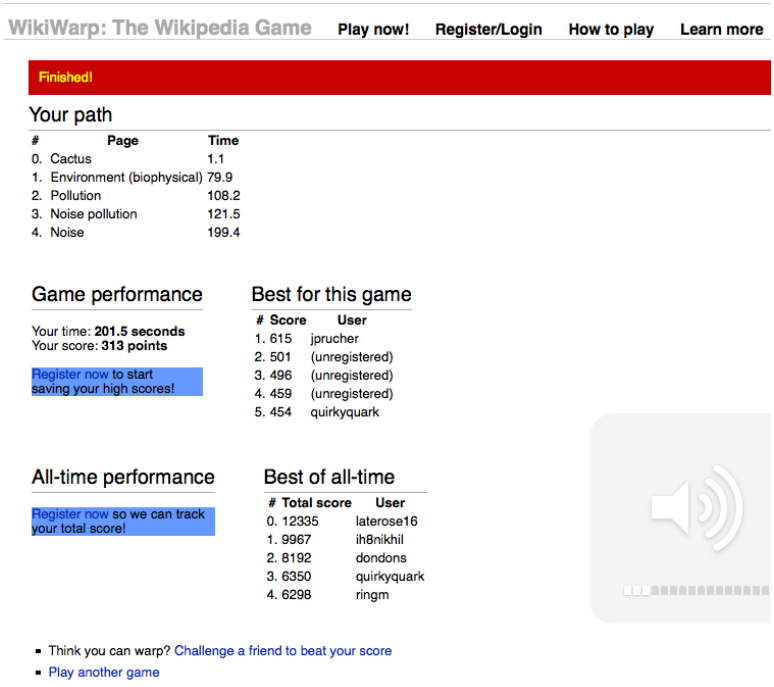

Figure 3: A completed game.

Back-end databases store the successful "warps", each of which represents one solution path to a particular featured game. Each custom game and warp can also be associated with a registered user who created it, and the website allows and encourages account creation to view and compare one's high scores and cumulative total score with others. (Users are also encouraged to invite and challenge their friends to beat their scores in particular games.) Each unique page is parsed only once and stored on the server for faster reloads.

4.1 Iterative Improvements and Important Lessons

The pilot study was launched on December 19, 2008 and kept online for two full weeks. Several inefficiencies in game design became evident soon after launch.

First, it was apparent that a sizable fraction of games started were ended incomplete, with users presumably frustrated by lack of progress or simply bored of the

current path. To accommodate, one third of the featured games thought to be most difficult were removed and the exponential decrease in the score as a function of time was slowed. A "Try a new game" button was considered, but rejected on the thought that it might dissuade players who were otherwise close to finishing. In the future, a "hint" feature might be considered (see Future Work). An important lesson here is the importance of real-time observation of every aspect of gameplay - achieved here by monitoring server logs, and in previous GWAPs like Tag-a-Tune conducting pencil and paper beta tests - to identify inefficiencies. In fact, later versions of the game will save incomplete warps in a separate database for further analysis instead of discarding the data.

Second, a rapid rise in popularity of the website - probably due to some attention from social bookmarking sites - caused dramatic slowing of server response times for all users. The problem could not be identified in real time (that is, without shutting down the site temporarily and losing users), but the error is expected to be a result of poor implementation and failing to test for scalability. Instead, the accumulated data was regularly saved and the server restarted to keep down loading times. The "invite your friends" feature was also disabled to keep traffic low. Nevertheless, for the second week of the study the website averaged 20 second loading times per page, making gameplay prohibitively slow.

Third, user statistics showed a minority (13%) of visitors to the site registered to record their scores and compete with others. This is an undesirable statistic, if only for the fact that valuable analysis can be conducted on performance variation between different user types, and having more user information would further this goal. However, low registration is not an uncommon phenomenon (particularly for a new website), and can be explained by modeling the registration process to have some associated cost that interferes with game enjoyment for the typical user.

To overcome this, users must be convinced that the benefits of registration - the thrill of inter-user competition, or possible prizes for high scorers - outweigh this cost. Prizes are an expensive but probably effective way of achieving this; a more sustainable solution would be to encourage a real sense of community within the website that itself encourages participants to join. Instead of playing against other usernames, users would compete with friends; this is indeed the circumstance in which the game was informally invented. Associating user profiles with social networking sites, for example, could build this community.

Finally, despite representing a small minority of total visitors, registered users accounted for a larger percentage (27%) of total completed games. Moreover, of this subset on a small fraction (17%) accumulated fully half of the completed games. This data is in accordance with informal observation; WikiWarp appeals to only a few individuals, but can be especially addictive for that group. In fact, over the two-week pilot study five individuals accumulated more than 45 minutes of playing time each. As with other peer production platforms such as Wikipedia itself or Yahoo Answers, consideration might be given to encourage the high-volume content creation from these "expert" users and to analyze their performance in comparison with others.

5 Results - Enjoyability

Maximizing the enjoyability of a GWAP is of primary practical concern to increase user participation and the volume of data generation.

Over the two week period the pilot study was active, over 800 games were played by more than 400 users, generating 50+ hours of gameplay. This is especially encouraging considering a large majority of the traffic occurred in the few days before the unexplained slowdown. Figure 4 is a chart of games played over time overlaid by the number of emails received over the same period complaining about site slowness.

Even more encouraging is the prevalence of "expert" users, as mentioned above, who accounted for a disproportionately large number of completed games in the pilot study. Three of the earliest slowness complaints came from this user group, whose content creation was interrupted by the site malfunction. It is possible to remain optimistic that after the implementation problems are solved, high user traffic and

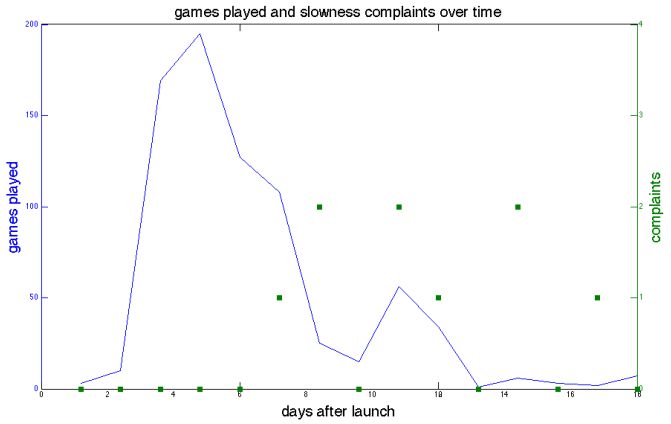

Figure 4: Completed games and slowness complaints over the two-week pilot study.

growth can be possible with consistent data generation from high-volume users.

6 Results - Usefulness

Enjoyability is a necessary but insufficient condition for a successful GWAP - equally as important is the requirement that the collected data be useful for the desired computational purpose. As such, data validation is essential even at the preliminary stages of a GWAP where data may be sparse.

6.1 What to Validate

It is important to note that in most other GWAPs in existence, a feasible map between the collected data and a computationally useful application is assumed to exist. For example, for the ESP game it is assumed that once images are tagged with relevant keywords, this data can be useful for visual search algorithms that return images based on a text query. Figure 5 lists examples of this map for various GWAPs.

Game	Collected Data	Computationally Useful Application
ESP Game	keyword-labeled images	image retrieval
Squiql (Peekaboom)	key-word labeled regions	image retrieval
Tag-a-Tune	audio clips labeled with keywords	audio retrieval, audio CAPTCHAs
Verbosity	common sense facts	natural language processing, AI
WikiWarp	*directed paths through Wikipedia*	information retrieval, etc.

Figure 5: Relationship of collected data to useful output for popular GWAPs.

Importantly, it is often impractical to validate the data by constructing a full-blown information retrieval application and analyzing its effectiveness - instead, researchers focus on verifying that the collected data seems to be "correct". Since there is no automated way to perform this verification (as this verification would itself represent a solution to a computationally intractable problem), data validation is usually performed directly by humans on subsets of the collected data. For example, the data validity of Verbosity is asserted by subjecting 200 of its random generated facts to human evaluation [5], and a paper introducing Tag-a-Tune presents example audio labels that *seem* accurate to claim validity [4].

For this problem of semantic relatedness, however, the most basic representation of the data is less suitable for direct examination by humans. One can compare paths by direct inspection - the following path:

- Prison(0.4), Kitchen (132.6), Eating(142.3), Food(148.6), Bread(156.6), Sourdough(171.2)

seems to be more meaningful than this one:

- Prison(0.2), Jerusalem(49.5), Judaism(97.9), Christianity(107.2), Holy Communion(149.6), Last Supper(184.3), Leaven(332.3), Dough(346.0), Bread(385.1), Sourdough(413.4)

but it is hard to quantify this agreement. Moreover it is the collection of sufficiently many paths in the aggregate that are necessary for even the most basic applications, and this assembly may reveal insights and subtleties unobservable from the individual pieces of data. Thus different metrics are required for determining the validity of semantic relatedness data.

6.2 How to Validate

In reality, there seem to be as many metrics of semantic accuracy as there are research teams in the field. Several methodologies, however, are more common than others. One is constructing an ad-hoc search algorithm generated from the generated data and presenting it with queries, such as "jaguar car" and "jaguar jungle", intended to draw out semantic distinctions that are difficult for computers to make. Another is the simple presentation of the most meaningful words or concepts by rank with respect to a given word, such as a list of the top 10 most relevant Wikipedia pages to "jaguar". With suitable methods of natural language processing, this process can be applied to text of arbitrary length. From a paper using statistical analysis on the texts of full-length Wikipedia articles come these results that demonstrate the power of both metrics:

Ambiguous word: "Jaguar"	
"Jaguar car models"	*"Jaguar (Panthera onca)"*
Jaguar (car)	Jaguar
Jaguar S-Type	Felidae
Jaguar X-type	Black panther
Jaguar E-Type	Leopard
Jaguar XJ	Puma
Daimler	Tiger
British Leyland Motor Corporation	Panthera hybrid
Luxury vehicles	Cave lion
V8 engine	American lion
Jaguar Racing	Kinkajou

Input: *"U.S. intelligence cannot say conclusively that Saddam Hussein has weapons of mass destruction, an information gap that is complicating White House efforts to build support for an attack on Saddam's Iraqi regime. The CIA has advised top administration officials to assume that Iraq has some weapons of mass destruction. But the agency has not given President Bush a "smoking gun," according to U.S. intelligence and administration officials."*
Iraq disarmament crisis
Yellowcake forgery
Senate Report of Pre-war Intelligence on Iraq
Iraq and weapons of mass destruction
Iraq Survey Group
September Dossier

Figure 6: Most-relevant concepts to presented search strings or natural language texts.

Note that both of these methods rely on the common sense of the reader to establish whether the system is performing properly.

Unfortunately, the data collected from WikiWarp is too sparse to perform a reliable disambiguation task between related concepts. (In retrospect, it may have been better to fully graph a subset of the articles in the online encyclopedia than constructing paths for unrelated sets of articles.) But it can still be instructive to view the most-relevant set of Wikipedia pages to a given page, assumed to be a proxy for the semantic relatedness of the two concepts, as determined by the collected data. To this end, a graph was generated between all visited pages in the full set of 800 games, with the value of the edge between pages i and j given by:

$$a_{ij} = \sum_{allpaths} f_{ij}(d, \bar{d}, t, \bar{t})$$

where f_{ij} was a function that was nonzero only if both pages i and j were visited by a given path, with its value some function of the distance d in links between i and j in the given path, the average number of links $\bar{d}$ over all paths for that featured game, the distance in time t between i and j in the given path, and the average total time $\bar{t}$ for

that featured game. Consideration was given to more advanced functions, but a simple intuitive function was ultimately chosen for at least preliminary data evaluation:

$$f_{ij} = exp(1 - d/\bar{d} - t/\bar{t})$$

This is high for adjacently-visited pages and pages visited in quick succession, and dies off quickly for pages separated in links or time. Using this metric, the "query results" for sample topics is presented below as an ordered vector of the five most closely related pages from the a_{ij} graph:

- **Guillotine**: Decapitation (5.34), Reign of Terror (1.03), French Revolution (0.68), France (0.09)
- **Bathing**: Hygiene (13.81), Bathtub (2.55), Bathtub hoax (0.48), Ceramic (0.02)
- **Dialect**: Dialect continuum (3.96), Language (3.37), Dialect (2.66), Australian Aboriginal languages (1.05), Variety (linguistics) (0.82)
- **Absolute zero**: Temperature (9.85), Celsius (3.93), Instrumental temperature record (2.62), Absolute zero (1.32), Absolute temperature (0.88)
- **Chemistry**: Organic chemistry (10.24), Fields of science (2.84), Geochemistry (1.43), Physical chemistry (1.25), Chemical potential (1.23)
- **Sound**: Noise (5.15), Hum (sound) (1.84), Acoustic location (1.81), Acoustics (1.05), Mantra (0.9)

- *Metal*: Coin (7.1), Brass (3.09), Purse (1.3), Brass instrument (0.86), Metalloid (0.8)
- *HIV*: Virus (5.2), Blood-borne disease (4.63), Syringe (0.86), Lentivirus (0.24)
- *Evolution*: Evolution (15.52), Human evolution (5.23), Evolutionary origin of religions (0.86), Evolutionary biology (0.78), Life Sciences (0.61)
- *Global warming*: Instrumental temperature record (4.98), Global warming (2.84), Temperature (1.39), Celsius (0.56), Absolute zero (0.32)

- Road junction: Road (5.66), Stop sign (4.55), Road number (0.7), Interchange (road) (0.51), Interstate Highway System (0.36)
- Noise: Noise pollution (5.87), Sound (5.15), Noise (environmental) (1.64), Noise health effects (1.3)
- Brass: Brass instrument (6.46), Metal (3.09), Alloy (1.39), Coin (0.62), Bronze (0.3)
- Temperature: Instrumental temperature record (18.31), Absolute zero (9.85), Temperature (4.16), Global warming (1.39), Fahrenheit (1.04)
- Democracy: Voting (3.48), Liberal democracy (2.49), Voting system (1.28), E-democracy (0.96)
- Growling: Dog (3.04), cat (1.66)
- Hygiene: Bathing (13.81), Shower (0.68), Bathtub (0.35), Disinfection (0.14), Showers (0.14)

Results are obviously better in sections of the graph that received more page visits. They also appear to be stronger for target pages (in bold) rather than start pages (italicized). This is consistent with the informal observation that users do not necessarily identify a semantic path and then follow it as closely as possible using Wikipedia links, but rather identify at each stage the link that will move them closest to their target. The first jump, for these games, is often quite drastic and has less semantic relevance. (This opens up the possibility of weighting by link number). Finally, the query vectors for the most-visited pages are relatively robust across changes in f_{ij}, indicating the results may be approaching some definitive measure of the semantic relatedness between concepts.

6.3 Future Work

6.3.1 Gameplay

In addition to the considerations listed in "Iterative Improvements and Important Lessons" - namely: fixing speed problems, increasing scalability, developing a stronger user network, and encouraging expert users - a number of improvements can be made to the existing gameplay that would result in increased enjoyability.

First, the difficulty problem can be addressed by staging beta tests of new featured games to obtain a baseline difficulty rating for a given game. This can be used to eliminate extremely easy or hard games, and even to modulate score based on difficulty. The past performance or preference of a player could change the difficulty of games he or she encounters. Finally, a system of hints could be put in place to direct a player toward their target, perhaps by listing the set of pages that link to the target page or the set of pages commonly used by other players for the same task, with an appropriate reduction of score.

Second, variations might be considered that increase human-human interaction in gameplay, such as allowing users to navigate toward any common page from random start pages or to race each other in completing games. Such competition might enable multiplayer game-theoretical analysis similar to that developed for other GWAPs. [2]

6.4 Data

One immediate exploration of the current data set would be exploring the dependence of the human-assessed accuracy of the most-related page vectors on the parameters of the f_{ij} function, and considering new parameters for such a function. This exploration might even utilize more advanced graph theory techniques to obtain a better closest-neighbor estimate (other than simply comparing edge strengths).

As mentioned previously, it might be fruitful to restrict gameplay to a subset of Wikipedia pages in the hopes of creating a denser graph structure - this limitation, however, may have adverse effects on the freedom of gameplay. Finally, as an alternate application and validation method, the data generated by WikiWarp could be used to create an intra-Wikipedia search engine, as the current search functionality is almost entirely text-matching and quite poor overall.

7 Conclusion

In this paper I have presented WikiWarp, an online GWAP that serves as a computational system to extract semantic relatedness between concepts by aggregating paths through Wikipedia's link graph. The game's motivation, design and implementation have been described, and some preliminary results have been demonstrated of its enjoyability and usefulness.

Results from the two-week pilot study are encouraging: despite less than optimal implementation, traffic and output was significant; despite sparse data, early validation checks are promising. It is hoped that this project, with continuous maintenance, can serve as a novel implementation of human computation and can continue to provide valuable data on semantic relatedness.

8 Acknowledgements

I am indebted to James Somers at the University of Michigan [jsomers.net] for algorithms to parse Wikipedia pages, much of the front-end design, and consistently good development advice.

9 Bibliography

1. von Ahn, Luis and Deabbish, Laura. *Labeling images with a computing game.* Conference on Human Factors in Computing Systems, 2004.

2. von Ahn, Luis and Dabbish, Laura. *Designing Games With A Purpose.* Communications of the ACM, August 51-8 (2008).

3. Gabrilovich, Evgeniy and Markovitch, Shaul, "Computing Semantic Relatedness using Wikipedia-based Explicity Semantic Analysis".

4. L.M. Law, Luis von Ahn et al. *Tagatune: A Game for Music and Sound Annotation.* Austrian Computer Society (OCG) (2007).

5. von Ahn, Luis et al. *Verbosity: A Game for Collecting Common-Sense Facts* CHI 2006 Proceedings (2006).