Field Research for Humanities with Social Media: Crowdsourcing and Algorithmic Data Analysis

François Bry, Fabian Kneißl, and Christoph Wieser

Ludwig-Maximilians-Universität München
Institut für Informatik
{bry|fabian.kneissl|christoph.wieser}@ifi.lmu.de

Abstract: Humanities rely on both field research data and databases but rarely have the means necessary for employing them. Crowdsourcing on the Web using social media specifically designed for the purpose offers a promising alternative. This article reports about two endeavors of this kind: enriching an art history database with semantic interpretations and collecting and assessing the regional and social origins of parts of speech for a linguistic investigation. The article motivates and describes the approach and further introduces into the semantic analysis method based on higher-order singular value decomposition specially designed for the project.

1 Introduction

In our daily life, Web search engines such as Google, Yahoo!, and Bing have become ubiquitous. Retrieving data in a Web of currently more than 130 million websites[1] without Web search engines would be hard to imagine. Current state-of-the-art Web search engines are permanently performing three tasks: collecting data (crawling), analyzing data (indexing), and returning answers. Crawling is the process of combing the Web for data. Indexing prepares collected data for returning answers. Finally, returning answers is a process triggered by customers of Web search engines: If a customer enters a query, search results are returned immediately.

Web search engines do a marvelous job collecting data available on the Web. However, data that is not digitally available is not amenable for Web crawlers. For example, data collected in interviews conducted for field studies, like those performed in linguistics, or data gained as taggings of artworks cannot be discovered directly by Web crawlers. One barrier for Web crawlers in the interview example (besides digitalization) is providing incentives for humans to reveal information. Another barrier for Web crawlers in the tagging example is interpretation, e.g., of an artwork. Both barriers can be broken down rather easily by playing humans: They can be "crowdsourced".

In humanities, there is an immense need for empirical data such as descriptions of artworks in art history or peculiarities of languages in linguistics. Such data is usually not digitally

[1] http://www.domaintools.com/internet-statistics/

available on the Web and in many cases there are no algorithms available to provide such data.

We let humans and computers do what they do best. Humans are able to assign semantic tags to content. For example, a human can recognize quite easily whether an artwork belongs to the middle age, but there still is a long way to go before algorithms capable of the same task can be devised. In contrast, computers can engage massive computational power. The key for sustaining the separation of concerns between what humans are good at and what computers are capable of lies in providing incentives for the humans to contribute in the human-computer joint task. In the project this article reports about, we harness the play instinct and curiosity of humans.

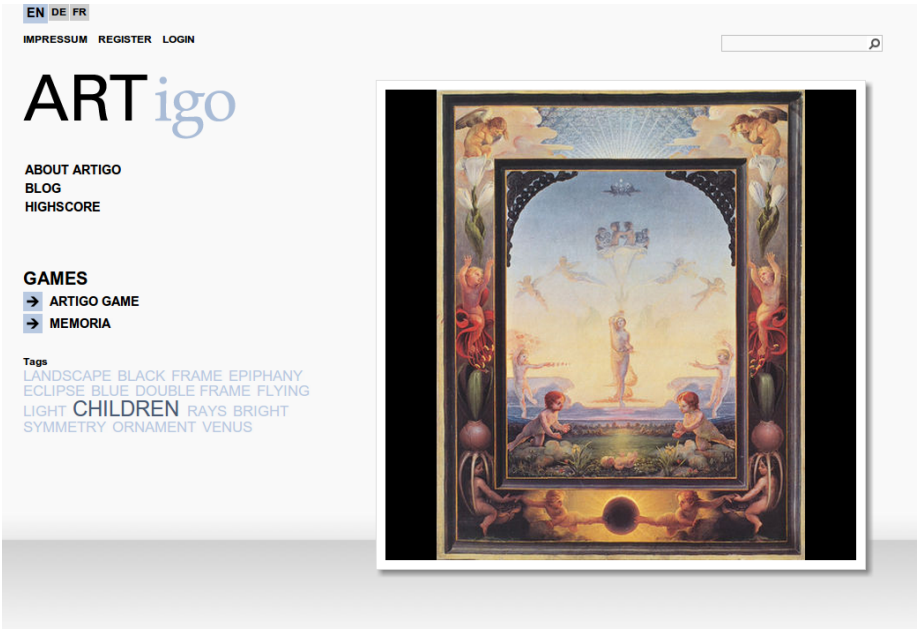

Figure 1: ARTIGO.

We developed two families of games: ARTIGO and METROPOLITALIA. ARTIGO[2] is based on the concepts of an image labeler. Two players, who only can communicate with each other through the tags they assign to a picture, try to find tags for a picture – in our case an artwork. Whenever two tags of the players match, each player is rewarded with points. The goal of ARTIGO is to gain meaningful semantic description of artworks so as to provide a powerful search engine. METROPOLITALIA deals with the peculiarities of the Italian language. As a kind of online field study, linguistic peculiarities from various parts of Italy are being collected. The key incentive for playing METROPOLITALIA is curiosity in the mutual perception of humans including self-reflection as well as the widespread interest in language among humans.

[2]This game platform is available at `http://www.artigo.org`.

The game families ARTIGO and METROPOLITALIA have in common that humans both provide information and profit from the collective input they provide. In other words, these game families launch an ecosystem. Thereby, the profit for users is twofold: Macroscopically, humans can exploit the knowledge of their predecessors that provided interpretations according to the current game design by using the platform's search feature. Microscopically, humans build up an individual profile by providing personal interpretations. Their input serves as basis for personal search results. Statistical methods permit to couple the microscopic view of a user to the macroscopic view of the social community of all users. This coupling results in a sort of individual filtering of all possible search results based on personal contributions. As mentioned above, the game design is crucial for the kind of information that can be harvested and exploited in the aforementioned ecosystem.

The contribution of this article lies in describing an approach towards improving database search and field research in humanities. Field research is improved because game families such as ARTIGO and METROPOLITALIA are (1) much cheaper than traditional field research, (2) the data collected by this form of crowdsourcing is more objective than that collected with traditional methods, and (3) a wide and widespread audience can be reached on the Web.

Search is improved because we can draw on individual perceptions of humans. On the one hand we can return answers in their own words and on the other hand we can couple the various perceptions of all users to generate semantically very rich answers.

The structure of this article is oriented towards the tasks of current search-engines: collecting data, analyzing data and returning answers. The first section is this introduction. The second section is devoted to related work. The third section discusses collecting data. The fourth section addresses data analysis. A fifth and final section is a conclusion.

2 Related Work

As of field research with social media, *Games With A Purpose* (GWAPs) yield valuable input for designing new games. Von Ahn and Dabbish first introduced this term for labeling images [vAD04]. Here, the so-called *ESP game* is presented in which the same image is shown to two randomly paired persons who type in words until they both come up with the same word. As the only commonly shared resource is the image, the players tend to enter descriptions of the image that are probable to be given also by the partner. A set of so-called taboo words for each image prevents common words to be entered by the players. Verified tags automatically convert to taboo words so that known description tags of an image are prevented from being entered again. This and an enforced time limit add a challenge aspect to the game.

In general, games with a purpose can solve a variety of different goals. Therefore, many games targeted at a special task have been proposed, e.g., games for protein-folding or eliciting user preferences (see [CTB+10] resp. [HvA09]). Not only games with a purpose gather data from the crowd, but also knowledge base systems like Wikipedia. [DRH11] provide an analysis and categorization of such systems. Furthermore, [DRH11] gives a

good and up-to-date overview of crowdsourcing Web platforms.

One further example where humans implicitly share their cognition skills in recognizing images is reCAPTCHA [vAMM+08]. *Completely Automated Public Turing tests to tell Computers and Humans Apart* (CAPTCHAs) are employed to prevent automated programs to act as humans and, e.g., create email accounts at free email service providers for sending spam. [vAMM+08] display an image which contains distorted characters scanned from books to users. The image has to be deciphered by humans and entered as text in order to pass the CAPTCHA. With the entered text, the recognition rate for *Optical Character Recognition* (OCR) software can be improved by displaying scanned words which were not recognized by this software. A neat side effect is that if an algorithm solving this problem was found for bypassing the CAPTCHAs, OCR software is improved and thus is a success for the project as well.

In spite of the assumption of the ESP game being solvable only by humans, for the image labeling game with taboo words [WRV09] developed an algorithm – ignoring the image! – which yields a good success rate of 69% for all images and 81% for images fulfilling one often-found property [WRV09]. Many times, synonyms of displayed taboo words are entered by human players and thus an algorithm can easily be devised that guesses which synonyms could fit to the image.

From a theoretical point of view, the game theory behind games with a purpose is investigated in [JP08] which explore how incentives change the equilibrium of a game and thus the results achieved. [HC09] abstracts from a specific game and proposes formal models for verification. Considerations in this work are applied to the implementation of ARTIGO, whereas the more general view of perception for field research proposed in section 3.2 is not considered and cannot be captured by the model proposed in [HC09].

The verification strategies for asserting that a user-given input can be accepted as a correct solution, however, are limited if no computer algorithm can verify its correctness. In [vAD08], three game-structure templates for two-player games are summarized: output-agreement games, inversion-problem games and input-agreement games. Most existing games with a purpose are based on one of these principles, as some kind of verification needs to be in place. Also the basic game of the ARTIGO game family is a typical output-agreement game. Indeed the tags entered by a player are verified through multiple players (who have to enter the same tags).

3 Collecting Data by Human Computation

3.1 ARTIGO – Artwork Tagging with General and Distinguishing Tags

ARTIGO is an online platform for games around crowdsourced search for artworks like paintings and sculptures. Especially in the humanities, art resources are inadequately made accessible for search. Two widely used German image databases *Bildarchiv Foto Marburg* and *Prometheus* both lack capabilities for searching for features of an artwork except title, artist and location. Also in general, it is difficult to provide a well-searchable image

Figure 2: ARTIGO Game.

database as no algorithm is known till now that recognizes images. Being both a gaming platform that employs human computation as well as a search engine, ARTIGO remedies this current problem.

The ARTIGO game itself (see figure 2 for a screenshot of an exemplary game session) is developed so as to be as easy to play as possible and at the same time to be sophisticated enough to gather the data sought for. A game consists of 5 rounds each lasting 60 seconds (which is a good value for enjoying game play according to [vA07]) and in each round one artwork is presented to the player who may enter tags for this shown piece of art. The number of the partner's tags is displayed as well as the player's current points and the entered tags. Furthermore, the tags that have been given by both players (direct matches) and the tags that have ever been given to the image at least once (indirect matches) are highlighted to provide the player with feedback about his performance. More points are awarded to direct matches than indirect matches to highlight the feeling of achievement with a human partner. If no human partner is available when a user wants to play, the server replays a previously recorded round to simulate a player's behavior. At the end of the game, a summary of the session is shown where players can review the artworks, additional meta data like author and title, and their entered tags.

As mentioned in the introduction, the first step towards searching consists in collecting information. The goal of ARTIGO in this process is to gather tags descriptive for an artwork. Two types of tags are necessary to get a broad range of tags: *general* as well as *distinguishing* tags.

A kind of image labeling game as proposed in [vAD04] is good for collecting general

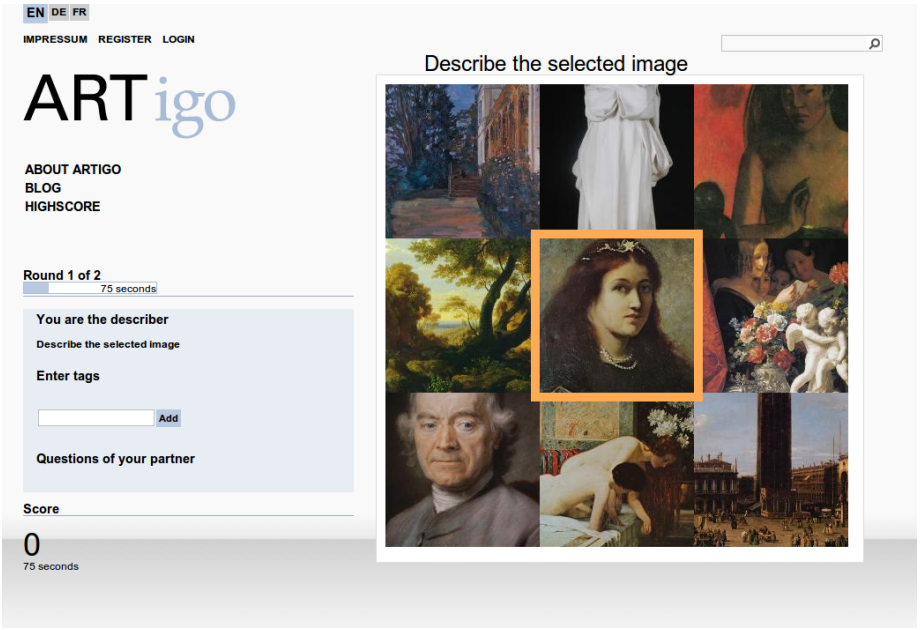

Figure 3: ARTIGO – Memoria game.

tags that describe the contents or characteristics of an artwork. To get more specific tags, a set of taboo words can be constructed automatically for each artwork as described in section 2. However this can limit the imaginativeness of players. So as to collect more specific tags, we therefore also propose a game without taboo words thus allowing some tag to be repeatedly given to a same artwork. This furthermore has the advantage that a frequency distribution of tags is generated for each artwork.

In order to retrieve distinguishing tags, further games called "Memoria" (see figure 3) are also provided on the ARTIGO platform. In such games, a player describes one out of a set of nine artworks. His partner must guess from the description which artwork is described. The difficulty of the game and thus how tags differentiate artworks can be adjusted by choosing artworks differently, for example, artworks close to each other (due to a high number of common tags) are more difficult to distinguish.

3.2 METROPOLITALIA – Italian Language Perception in Major Cities

METROPOLITALIA introduces a new element in terms of games with a purpose: The *language itself* is the subject of the games. Particularly in today's Italian, new standard forms establish themselves in large cities. The language is rapidly changing and nobody knows exactly how (and why). Employing traditional field research methods is hardly feasible because of the high effort and therefore high costs and because of the subjective percep-

tion of the participating field researchers who, most likely, would bias the data collected. Here, crowdsourcing with games with a purpose provides a cheap and continuous way of gathering data from a wide and widespread community.

Several games are currently being implemented to explore the above-mentioned evolution of the Italian language. One of them – we take it as an example in this article – consists in native Italian speakers guessing where in Italy a given statement is spoken. The statement can either be spoken (presented as an audio file) or written and consists of a (part of a) sentence typical for a region or city in Italy. After guessing, the player's assignment of a location to a statement is compared to the other players' assignments and shown to him to provide appropriate feedback for self-reflection and to increase the joy of the game. Like all people, Italians are generally very fond of their language and like to learn about it. We believe that the enjoyment to experience differences in one's own language provides enough incentive for a large number of Italian speakers to play on the platform METROPOLITALIA, once the platform is sufficiently known. The data acquired through METROPOLITALIA in the suggested game consist of triples user–statement–statement location or quadruples user–user location–statement–statement location and therefore give rise to promising data analyses.

In contrast to the games with a purpose mentioned in section 2, *field research* often deals with the acquisition of people's perception differences and thus a definitive answer to a problem is not possible and furthermore not desired. In METROPOLITALIA, for example, players assign locations to statements from local speakers, thus defining the statements' distribution. Possible results consist in all players assigning the same town or – in contrast – every player assigning a different location. Both results can prove useful for research and it is therefore difficult to establish a verification schema. However, one can argue that also the described METROPOLITALIA is a kind of output-agreement game in that players assign a location as output and if most players assign the same location, this must be the right one. For certain games, this argumentation is correct, but in general a different approach has to be taken for field research.

Therefore, we suggest to not use a scheme of verification, but of *perception* and weaken the requirement for the answer being correct (as far as "correct" is a possible output). Two types of perception become apparent:

- *Self-perception* represents the difference in between a single player and all other people. This enables the player to assess how he performs compared to other players in form of high score tables or other comparative game play motivators.
- *External perception* captures the characteristics of a group of people, in the simplest case all players. Here, the output of this group can be – as outlined above – the mutual consent on one solution or a distribution of the perception to several opinions. Both cases are relevant for field research and complement the self-perception, which spurs the players' motivation, in form of an incentive for researchers.

The goal of METROPOLITALIA is to gather data for these two types of perception in the Italian language and provide the results to researchers as well as users to generate a dictionary-like database for statements and their perceived locations.

4 Data Analysis: Latent Semantic Analysis for Indexing

In the previous sections we focused on the (mainly) human part, i.e. interpreting data. This section focuses on the analysis of collected data performed by computers. As mentioned in the introduction, microscopically humans build up an individual profile by providing personal interpretations. Their input serves as basis for personal search results. We couple the microscopic view of each user to the macroscopic view of the social community of all users. That means, we need to take into account the different ways of tagging on the one hand and on the other hand we need to differentiate between worthy contributions and useless noise. The result is an improved search that returns answers in the own words of users and that generates semantically very rich answers.

We make use of individual perceptions by reducing the statistical noise produced by human users to crystallize out a latent semantic of all input data. The technique is inspired by Latent Semantic Analysis/Indexing (LSA) [DDL$^+$90][Dum04]. With LSA a collection of documents can be analysed according to occurrences of terms in documents. On basis mainly of Singular Value Decomposition (SVD) [Hog07], assignments of a tag to a document can be sharpened by reducing statistical noise. LSA reduces the noise of all tags[3].

However, LSA does not differentiate between different users. The input for LSA is a document-term matrix where all taggings are "anonymously" accumulated in the components of the matrix. From the perspective of a matrix component, one matrix component represents the number of humans that assigned a specific tag to one single document. For additionally representing the creator of a tag, an extension of the document-term matrix as data structure is needed. A natural extension to differentiate between users is to introduce one matrix component per user. This extension leads us from a document-term matrix to a document-term-user tensor.

A tensor [KB09] is a generalization of a matrix. For example, a 3rd-order tensor is a "3-dimensional matrix" consisting of matrices stacked one above the other. A matrix can be seen as a 2nd-order tensor. For representing documents, terms, and users a 3rd-order tensor is needed. The value 1 (or 0) of a component in a document-term-user tensor determines, if a user tagged a document with a term (or not).

The generalization of the matrix data structure to tensors entails a generalization of LSA as well. The main difference is the generalization of SVD (as main part of LSA) to Higher-Order SVD (HOSVD), that allows tensors as input and not only matrices. HOSVD allows to reduce the statistical noise of documents, tags, and users and not only one of them. The generalized noise reduction allows us to bring down all users to a common denominator, i.e., exploiting the *knowledge* of all users. The term *knowledge* refers to statistically significant taggings from the perspective of the community of users without statistical noise.

The result of a HOSVD of a document-term-user tensor with reduced noise allows to calculate individual and semantically rich answers in the language of a given user. The HOSVD as basis for calculating answers needs to be calculated only once for a given document-term-user tensor.

[3]The noise produced by documents is not reduced.

In first experiments we used a tensor with 100 terms, 100 documents, and 100 users containing taggings collected with ARTIGO. A comparison of LSA and higher-order LSA via 10-fold cross-validation for search strings consisting of 1–5 tags yielded better average precision results from 12.2% (matrix LSA) to 25.86% (higher-order LSA) and better average recall results from 4.5% (matrix LSA) to 9.9% (higher-order LSA). The decomposition of the tensor was done single-threaded in about 2.5 hours.

5 Conclusion

In this article, we described how social media, in particular games with a purpose, can be used for both enriching an image database used in art history as well as for linguistic field research.

The two applications are at different stages. The art history application has already gathered enough data for analysis. The linguistic application is still at its beginning.

Even though both applications are very different in the data they rely upon and collect and in their scientific objectives, they can rely on the same data analysis methods: Both can benefit from a higher-order, e.g., tensor instead of matrix-based, latent semantic analysis. The approach has been motivated and described.

Further steps in this crowdsourcing approach on field research for humanities include enhancing the platforms to a wider range of games, gather more data, and apply the suggested analysis methods to these data. Also, further means of verification are to be examined.

References

[CTB+10] Seth Cooper, Adrien Treuille, Janos Barbero, Andrew Leaver Fay, Kathleen Tuite, Firas Khatib, Alex Cho Snyder, Michael Beenen, David Salesin, David Baker, and Zoran Popović. The challenge of designing scientific discovery games. FDG '10, pages 40–47, Monterey, California, 2010. ACM.

[DDL+90] Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by Latent Semantic Analysis. *JASIS*, 41(6):391–407, 1990.

[DRH11] Anhai Doan, Raghu Ramakrishnan, and Alon Y. Halevy. Crowdsourcing systems on the World-Wide Web. *Communications of the ACM*, 54(4):86, April 2011.

[Dum04] Susan T. Dumais. Latent Semantic Analysis. *Annual Review of Information Science and Technology*, 38(1):188–230, 2004.

[HC09] Chien-Ju Ho and Kuan-Ta Chen. On formal models for social verification. In *ACM SIGKDD Workshop on Human Computation*, page 62. ACM Press, 2009.

[Hog07] Leslie Hogben. *Handbook of Linear Algebra*. Boca Raton CRC Press, 2007.

[HvA09] Severin Hacker and Luis von Ahn. Matchin: eliciting user preferences with an online game. CHI '09, pages 1207–1216, Boston, MA, USA, 2009. ACM.

[JP08] Shaili Jain and David Parkes. A Game-Theoretic Analysis of Games with a Purpose. *Internet and Network Economics*, 5385:342–350, 2008.

[KB09] Tamara G. Kolda and Brett W. Bader. Tensor Decompositions and Applications. *SIAM Review*, 51(3):455–500, September 2009.

[vA07] Luis von Ahn. Human computation. pages 5–6, Whistler, BC, Canada, 2007. ACM.

[vAD04] Luis von Ahn and Laura Dabbish. Labeling images with a computer game. In *SIGCHI conference on Human factors in computing systems*, pages 319–326, Vienna, Austria, 2004. ACM.

[vAD08] Luis von Ahn and Laura Dabbish. Designing games with a purpose. *Communications of the ACM*, 51(8):58–67, August 2008.

[vAMM+08] Luis von Ahn, Benjamin Maurer, Colin McMillen, David Abraham, and Manuel Blum. reCAPTCHA: Human-Based Character Recognition via Web Security Measures. *Science*, August 2008.

[WRV09] Ingmar Weber, Stephen Robertson, and Milan Vojnovic. Rethinking the ESP game. In *27th international conference extended abstracts on Human factors in computing systems - CHI EA '09*, page 3937, New York, New York, USA, 2009. ACM Press.