View of the world according to Wikipedia: are we all little Steinbergs?

S. E. Overell[a,*], S. Rüger[b]

[a] *Department of Computing, Imperial College London, UK*
[b] *Knowledge Media Institute, The Open University, UK*

Abstract

Saul Steinberg's most famous cartoon "View of the world from 9th Avenue" depicts the world as seen by self-absorbed New Yorkers (Figure 1). By analysing wikipediae of a range of different languages, we find that this particular fish-eye world view is ubiquitous and inherently part of human nature.

By measuring the skew in the distribution of locations in different languages we can confirm the validity of plausible quantitative models. These models demonstrate convincingly that people all have similar world views: "We are all little Steinbergs."

Our Steinberg hypothesis allows the world view of specific people to be more accurately modelled; this will allow greater understanding of a person's discourse, either by someone else or automatically by a computer.

Keywords: Social role identification, Modelling, Data Mining

1. Introduction

We hypothesise that people's view of the world is relative to their current location and that we all have the same way of thinking and reasoning about locations. Egenhofer and Mark's notion of Naive Geography is concerned with formal models of the common-sense geographic world (and how this differs from the physical world)[1]. We build on this notion quantifying the

*Corresponding author
Email address: seo01@doc.ic.ac.uk (S. E. Overell)

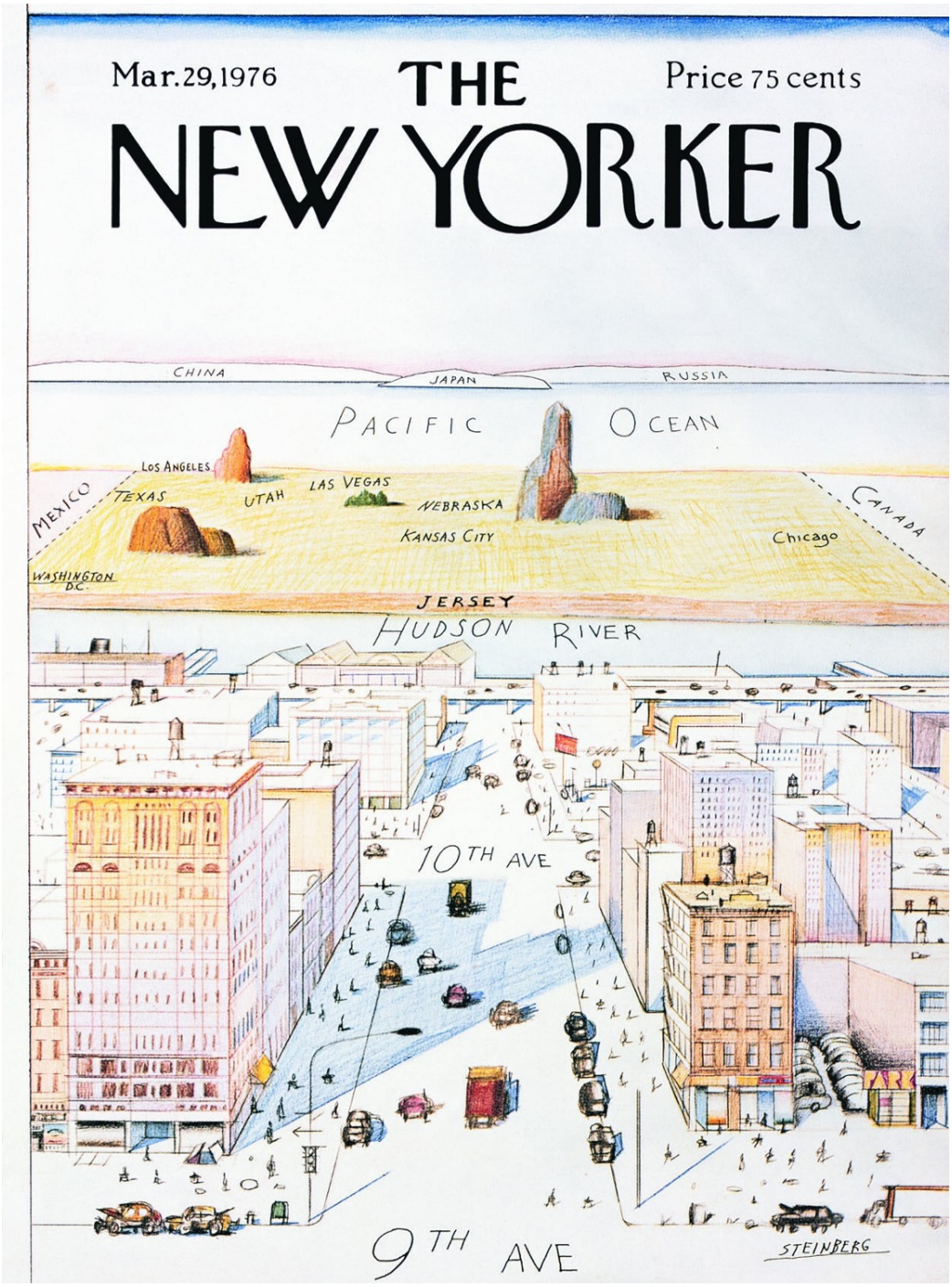

Figure 1: Cover of The New Yorker from March 29, 1976.

importance of a given location to a given person, based on their distance from the location and its population.

Over the past 10 years the proliferation of User Generated Content (UGC) has been substantial. Wikis, blogs and tweets are now ubiquitous and much of this data is available freely via APIs and bulk download. Similarly over this time the software and hardware required for data mining has been increasingly commoditised lowering the barrier to entry for researchers wishing to perform inference over this vast amount of content.

Recently we have seen a series of successful attempts to build simple predictive models using a small number of features to make broad socio-economic inferences off the back of this data. Notably Bollen et al's work predicting stock-market trends based on simple sentiment analysis of twitter feeds[2], Mishne and Glance's work inferring movie box office takings from blog analysis[3], and Azhar's recently launched web-site Peer-Index which infers influence from Twitter, Linked-in, Facebook, Quora, blogs and more[4].

Wikipedia is the largest reference web site on the Internet and typifies this growth in UGC[5]. Since its launch in 2001 it has grown to become the 8th most popular site on the Internet[6]. Its success is largely attributed to its wiki software, which allows any user to update almost any page at any time[7]. This keeps Wikipedia up to date (articles are often updated minutes after events happening) and extraordinarily extensive (in English: 3.6M articles to date with new articles created at a rate of 1,000 per day)[8]. Wikipedia has been described by its creator, Jimmy Wales, as "*an effort to create and distribute a free encyclopedia of the highest possible quality to every single person on the planet in their own language[9].*"

Wikipedia is now available in nearly 280 languages, with the English Wikipedia making up only one fifth of the total number of articles. Its extensive size, number of languages and popularity makes Wikipedia attractive to people wishing to mine information. Mihalcea recognised how Wikipedia's hyperlinked structure makes it ideal for generating annotated unambiguous corpora[10]. We exploit this structure and identify articles that refer to specific locations. By counting the links in the rest of Wikipedia to this classified set that we have generated, a distribution of references to locations can be built. It is this distribution of references to locations that we fit models against.

This article follows a methodology analogous to that used by Bollen et al., Mishne and Glance, and others of extracting simple features from user generated content to draw broad inferences and form predictive models[2, 3].

Section 2 describes our data-set of classified Wikipedia articles across six wikipediae of different languages. Comparisons are drawn between the data extracted for different wikipediae and we infer and visualise a world view of a *typical* speaker of a language. Section 3 presents and formalises our "Steinberg Hypothesis," that a skewed fish-eye world view is inherently part of human nature. Section 4 tests this hypothesis by estimating the likelihood of a person to refer to a specific location: the distance from a person to a location and the population of the location are taken as features, and references to classified Wikipedia articles as test data. We apply a series of decay and growth functions respectively to distance and population to find the optimal function and parameters. We conclude with Section 5 which outlines the novelty and contributions of this work, and presents its applications in computational science and beyond.

2. Classifying Wikipedia Articles

Classifying Wikipedia articles as locations is not a trivial task. In 2005 Wikipedia started a project to geotag their articles, embedding geographic co-ordinates linking to external location based services[11]. Further attempts to classify Wikipedia articles have compared the content of articles to an external ontology[12, 13], or used complex rules based on article content and meta-data[14, 15]. The approach used in this article, described in detail in Overell and Rüger (2007)[16], uses a pipeline of different classification techniques based on the article's title, categories, anchor texts and geotags. This pipeline matches articles in Wikipedia to the Getty Thesaurus of Geographical Names (TGN), an authoritative gazetteer of locations, with 94% accuracy.

Our pipeline to classify Wikipedia articles is language independent. In fact, we have applied it to six different wikipediae: English, Chinese, German, French, Portuguese and Spanish. For the non-English language versions of Wikipedia we add additional evidence, inter-language links to the English Wikipedia, to the classification pipeline.

These six sets of classified Wikipedia articles give us six location distributions (Figure 2). These location distributions clearly show the skew in the different wikipediae of which locations are discussed the most. Even in these visualisations our Steinberg hypothesis is apparent: locations where the respective language is spoken show a dramatic concentration of references. This skew can be seen in Table 1: for each Wikipedia we divide the world into

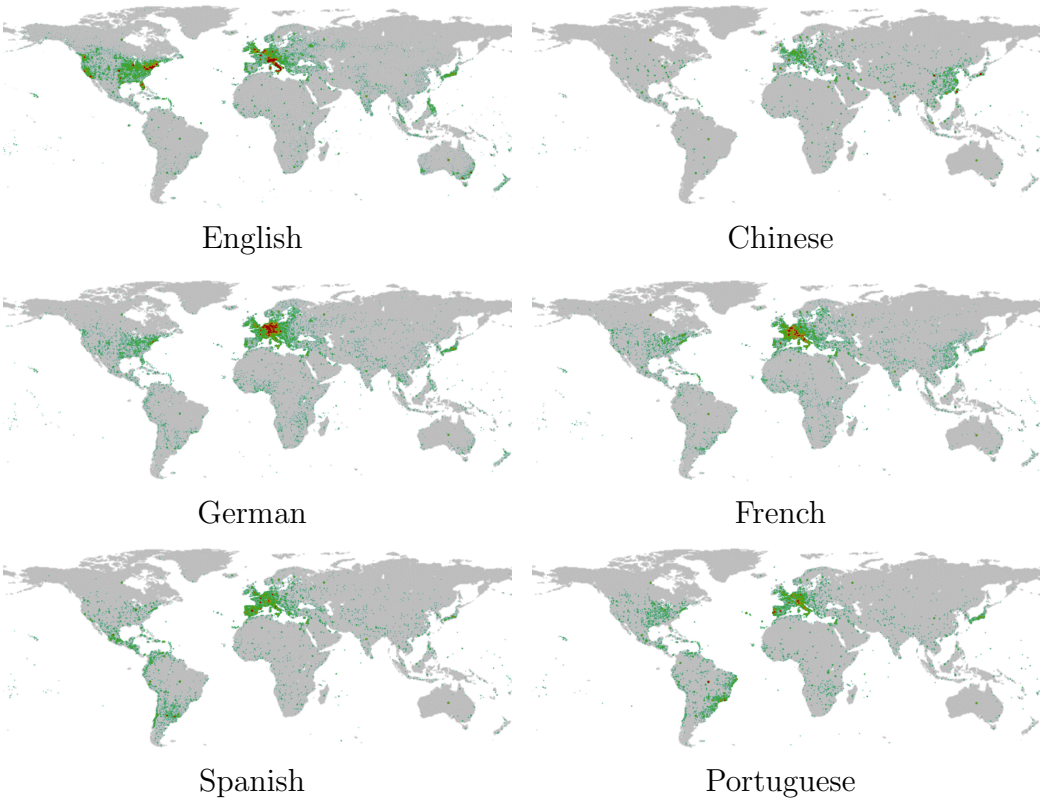

Figure 2: Heat maps of locations referred to in different Wikipediae. Links to articles describing locations are mined from each language version of Wikipedia and displayed as a series of normalised heat maps.

Table 1: References per 1M people to articles about locations in the different wikipediae.

Language	English	Chinese	German	French	Port.	Spanish
Refs. / 1M spkrs	2495	11.2	709	461	84	103
Refs. / 1M non-spkrs	189	14.2	24	23	12	25
Bias	13.15	0.79	28.97	19.88	6.90	4.09

two parts, locations where the respective language is spoken and locations where it is not. We then calculate the references per person in each division[1]. How biased a particular wikipedia is toward speakers of its language can be formalised as the ratio of these two numbers. For example, in the English language Wikipedia there are 2,495 references to places where English is spoken per 1M people that live there, and 189 references to places where English is not spoken per 1M people that live there. Notice Chinese is the only language with a bias < 1. We attribute this to the fact that Wikipedia has been blocked in China for significant periods of time, therefore many of the people editing Wikipedia in Chinese will be expatriates.

This skew can be visualised by deforming global maps so that the area of a country reflects its positive or negative bias. Figure 3 shows a series of "Steinberg Maps," i.e., maps deformed to the world view of a *typical* speaker of a particular language.

3. The Steinberg Hypothesis

Visualising the world as seen by all the speakers of a language is one thing. However the goal for this article is to model how a *single* person sees the world, in the same way as Steinberg does for a single New Yorker. We assume Wikipedia is read and edited by a typical sample of the population. By summing the predicted world views of a population and fitting this combined model to Wikipedia, we can test the validity of the individual models. To do this we define the relevance of a location to a person. We consider "relevance" in this context as a synonym for "likelihood to use in dialogue". We calculate the relevance of a location l to a person p as

$$\mathrm{rel}(l, p) = \mathrm{subjInt}(l, p) \cdot \mathrm{objInt}(l),$$

[1]Countries where multiple languages are spoken have references apportioned in the appropriate ratios.

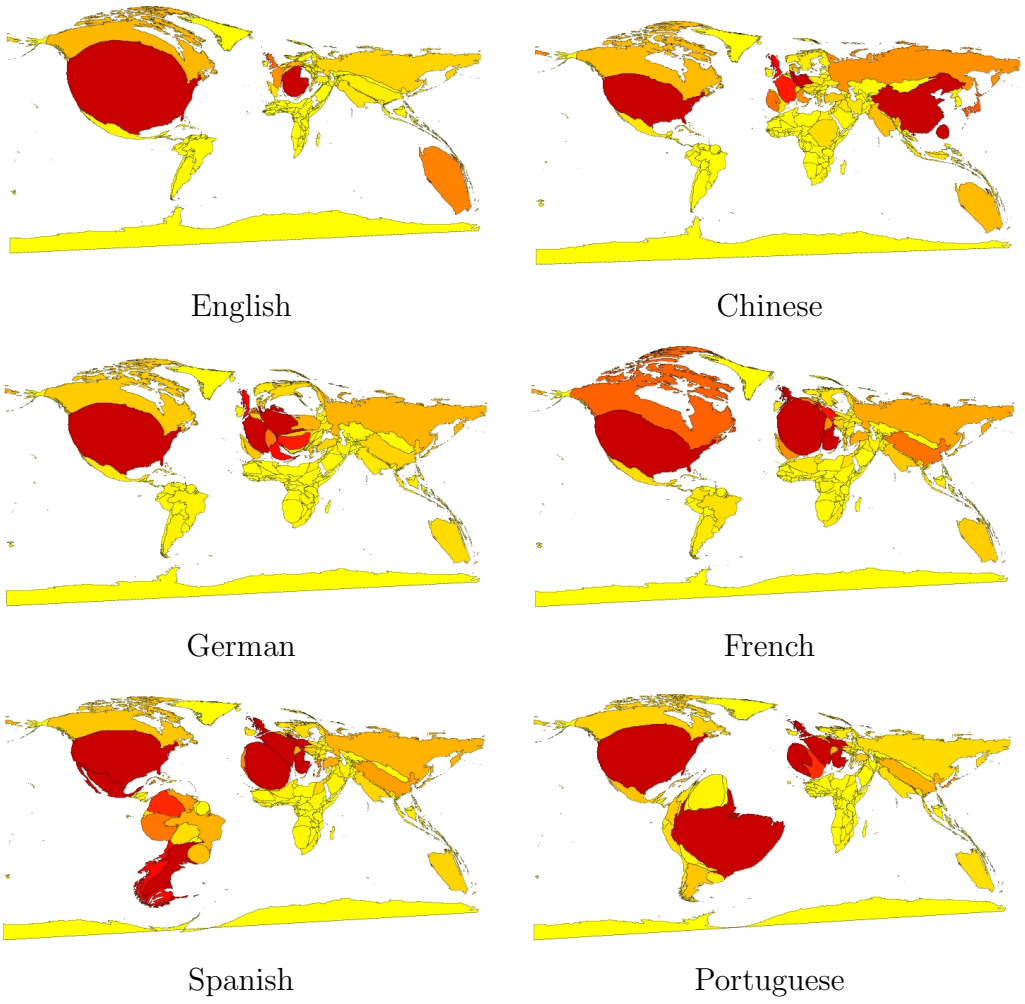

Figure 3: Steinberg Maps — The area of a country represents its positive or negative bias with respect to a particular Wikipedia (see Table 1 for how bias is calculated).

i.e., the product of the subjective interestingness of location l to person p and the objective interestingness of the location. The subjective interestingness is based on the relationship between p and l, while the objective interestingness is based on properties of l. We model the subjective interestingness as a function f of the distance from p to l:

$$\mathrm{rel}(l, p) = f(\mathrm{dist}(p, l)) \cdot \mathrm{objInt}(l).$$

In our experiment we compare different possible functions for f and $\mathrm{objInt}(l)$. When modelling objective interestingness the only property we will take into account is population. The distance function used is the geodesic distance, chosen for simplicity, subject to a minimum of 1 km to prevent asymptotic irregularities near zero. The relevance of a location l to person p can be seen as an estimate of the probability that p will refer to l in a document that s/he authors. We extend this to Wikipedia by averaging across all potential authors: let A_v be the set of all non-overlapping locations where v is spoken, $\mathrm{pop}(a)$ the population of location a, and $\mathrm{prop}(a, v)$ the proportion of people in location a that speak language v. Then our approximation of $\mathrm{rel}^v(l)$, the relevance of location l to all the speakers of language v is

$$\mathrm{rel}^v(l) = \sum_{a \in A_v} (\,\mathrm{pop}(a) \cdot \mathrm{prop}(a, v) \cdot f(\mathrm{dist}(a, l))) \cdot \mathrm{objInt}(l).$$

4. Results

To test our Steinberg hypothesis we compared different possible and plausible equations for $f(d)$ and $\mathrm{objInt}(l)$. Essentially we apply a series of simple decay and growth functions respectively to distance and population to find the optimal function and parameters. We consider the frequencies of links to Wikipedia articles describing locations a histogram and normalise these observed frequencies, O, to result in a unit histogram O'. A predicted histogram, P, is generated consisting of the set of predicted frequencies for all locations calculated using $\mathrm{rel}^v(l)$. P is normalised to give P'. To tune the variables in the formulations of $f(d)$ and $\mathrm{objInt}(l)$ we implemented an iterative greedy algorithm that minimises the symmetric difference between O' and P'. Our baseline assumed every location was equally likely to be referred to:

$$\mathrm{rel}^v(l) = 1$$

This was compared to three possible decay functions for $f(d)$:

$$f_1(d) = 1, \qquad f_2(d) = \frac{1}{\log_\beta(d)}, \qquad f_3(d) = d^{-\beta}$$

and two possible formulations for $\text{objInt}(l)$:

$$\text{objInt}_1(l) = \log_\alpha(\text{pop}(l)), \qquad \text{objInt}_2(l) = \text{pop}(l)^\alpha.$$

For each of the six languages the best fitting equation of the seven considered (3 $f(d)$ functions $\times$ 2 $\text{objInt}(l)$ functions + the baseline) turned out to be

$$\text{rel}(l, p) = \text{dist}(p, l)^{-\beta} \cdot \text{pop}(l)^\alpha.$$

Table 2 shows the total symmetric difference between O' and P' for the baseline and the best fitting case, and the optimal α and β values. In all cases the best fitting equation is statistically significantly better than the second best equation and the baseline, except with the Spanish Wikipedia where there is no significant difference between the top two methods. We attribute this to Spanish being the most widely spoken native language (geographically), hence may require a more sophisticated model. The significance test employed was a one tailed Student's t-test with a significance level of 2.5%[17]. The significant test was performed as a paired test across locations comparing the difference between O' and P' for each location. The higher the β values the lower the likelihood that far away places will be referenced, and the smaller the α values the higher likelihood that smaller places will be referenced. Notice both the α and β values are below one, meaning they have a sublinear relation with respect to population and distance, that is, exhibit a saturation effect. This means there is effectively a distance beyond which everything is considered *far away* and a population above which locations are considered *important*. As can be observed in Table 2, Chinese speaking editors of Wikipedia write more about distant places, while English speaking editors write more about smaller places. This concurs with the maps shown in Figure 2 and could be attributed to the fact that Wikipedia has been blocked in China for a significant periods of time, and the relative maturity of the English Wikipedia compared to the other languages.

5. Conclusions

The approach presented here is sufficiently accurate to support our Steinberg hypothesis: that *everyone* has a localised fish-eye view of the world. We

Table 2: The top part of the table shows the symmetric difference between the observed and expected number of references to locations in the baseline formulation of the equation and the best performing equation. The best performing equation for every language was $\mathrm{rel}(l,p) = \mathrm{dist}(p,l)^{-\beta} \cdot \mathrm{pop}(l)^{\alpha}$. This means the likelihood of a location to be referred to by a person can be modelled by the product of an exponential drop off of their distance from that location with parameter β and the exponential increase of the population of that location with parameter α. The symmetric difference is how much the model differs from the observations from Wikipedia. The optimal α and β values are shown in the lower part of the table.

Language	English	Chinese	German	French	Port.	Spanish
Baseline	1.05	1.15	1.17	1.23	1.19	1.23
Best	0.92	0.92	0.77	0.87	0.89	1.02
Optimal α	0.44	0.67	0.91	0.92	0.73	0.57
Optimal β	0.72	0.24	0.89	0.75	0.57	0.68

expect that including topographical distance, migration patterns, political, social and economic factors into the modelling process will achieve a more accurate predictive model. Models such as the ones presented here, despite being facile, can still have significant consequences when it comes to understanding a person's discourse. To provide a more concrete example, consider a geographically aware search engine with the ability to answer the query "Jobs in Cambridge." It is not apparent from the query whether Cambridge, UK, Cambridge, Massachusetts, or Cambridge, New Zealand, is intended. The approximate location of the user can be calculated from their IP address; Figure 4 shows, according to our model, which Cambridge is most likely to correspond to the user's intention based on their location, using the top performing function fitted to the English language Wikipedia.

We are unaware of any other studies quantifying the bias in the references to locations in a multi-lingual corpus as large and diverse as Wikipedia. The core contribution of this article is quantifying and validating the widely held belief that people have a localised world-view; and quantifying this difference across a variety of languages.

In a broader context recognising the phenomenon asserted by our Steinberg hypothesis could enrich human dialogue and increase understanding between people. On one level it demonstrates that bias and prejudice toward our own location are part of human nature and to some extent can be excused. On a higher level understanding this phenomenon can help avoid

Figure 4: Map of the world showing when someone refers to "Cambridge," which location they are most likely to mean dependant on where the are. If the person is in a green part of the world, they probably mean Cambridge Massachusetts, red is Cambridge, UK and blue, Cambridge New Zealand. This map is based on the best performing function with variables optimised for English speakers: $\mathrm{rel}(l,p) = \mathrm{dist}(p,l)^{-0.72} \cdot \mathrm{pop}(l)^{0.44}$.

confusion and increase the shared understanding of the world required for any dialogue.

References

[1] M. Egenhofer, D. Mark, Naive geography, in: Spatial Information Theory A Theoretical Basis for GIS, volume 988, Springer Berlin, 1995, pp. 1–15.

[2] J. Bollen, H. Mao, X. Zeng, Twitter mood predicts the stock market, Journal of Computational Science 2 (2011) 1–8.

[3] G. Mishne, N. Glance, Predicting movie sales from blogger sentiment, in: AAAI Symposium on Computational Approaches to Analysing Weblogs, pp. 155–158.

[4] A. Azhar, `http://www.peerindex.net/`, 2011. Accessed 1 January.

[5] J. Giles, Internet encyclopaedias go head to head, Nature 438 (2005) 900–901.

[6] Alexa Internet, Inc., http://www.alexa.com/, 2008. Accessed 1 October.

[7] D. Tapscott, A. D. Williams, Wikinomics, Atlantic Books, 2nd edition, 2008.

[8] E. Zachte, Wikipedia statistics, http://stats.wikimedia.org/EN/Sitemap.htm, 2011. Generated 31 March.

[9] The Wikimedia foundation, http://en.wikipedia.org/wiki/Wikipedia:Multilingual_coordination, 2008. Accessed 1 October.

[10] R. Mihalcea, Using wikipedia for automatic word sense disambiguation, in: Human Language Technologies, NAACL, 2007, pp. 196–203.

[11] The Wikimedia foundation, http://en.wikipedia.org/wiki/Wikipedia:WikiProject_Geographical_coordinates, 2008. Accessed 1 October.

[12] M. Ruiz-Casado, E. Alfonseca, P. Castells, Automatic assignment of Wikipedia encyclopedic entries to WordNet synsets, in: Advances in Web Intelligence, volume 3528, Springer Berlin, 2005, pp. 380–386.

[13] D. Buscaldi, P. Rosso, P. García, Inferring geographic ontologies from multiple resources for geographical information retrieval, in: SIGIR workshop on Geographic Information Retrieval, ACM, 2006, pp. 53–55.

[14] F. Suchanek, G. Kasneci, G. Weikem, YAGO: A core of semantic knowledge unifying WordNet and Wikipedia, in: WWW'07, ACM, 2007, pp. 697–706.

[15] S. Auer, J. Lehmann, What have Innsbruck and Leipzig in common?, in: The Semantic Web: Research and Applications, volume 4519, Springer Berlin, 2007, pp. 503–517.

[16] S. Overell, S. Rüger, Geographic co-occurrence as a tool for GIR, in: CIKM workshop on Geographic Information Retrieval, ACM, 2007, pp. 71–76.

[17] D. Hull, Using statistical testing in the evaluation of retrieval experiments, in: ACM SIGIR conference on Research and development in information retrieval, ACM, 1993, pp. 329–338.