A Probabilistic Model of Semantics in Social Information Foraging

Peter Pirolli

Palo Alto Research Center, Inc.
3333 Coyote Hill Road
Palo Alto, CA 94116
pirolli@parc.com

Abstract

Probabilistic theories of semantic category formation in cognitive psychology can be used to model how semantic sense may be built up in the aggregate by a network of cooperating users. A model based on Latent Dirichlet Allocation is applied to word activity data in the Lostpedia wiki to reveal the structure of topics being written about, the trends in attention devoted to topics, and relating the model inferences to events known to have occurred in the real world. Relevance of the models to progress in exploratory search and social information foraging is discussed.
.

INTRODUCTION

Over the past decade, the Web and the Internet have evolved to become much more social, including being much more about socially mediated information foraging and sensemaking. This is evident in the growth of blogs [10], wikis [20], and social tagging systems [16]. It is also evident in the growth of academic interest in socially mediated intelligence and the cooperative production of knowledge [2, 18, 19]. It has been argued [13] that models in cognitive psychology can lead to greater understanding and better designs for human-information interaction systems. With that motivation, this paper presents extensions of Information Foraging Theory [14] aimed at modeling the acquisition of semantic knowledge about some topic area at the social level. The model is an attempt to start to address the question: how do the semantic representations of a community engaged in social information foraging and sensemaking evolve over time?

Probabilistic Semantic Representations

Probabilistic representations—particularly those based on Bayesian approaches—of semantics have arisen as the result of rational analyses of cognition [1, 17]. Rational analysis is a framework in which it is heuristically assumed that human cognition approximates an optimal solution to the problems posed by the environment. Topic category judgments are viewed as prediction problems, and these can be guided by statistical inferences made from the structure of the environment, especially the linguistic environment.

The Topic Model based on Latent Dirichlet Allocation (LDA). Griffiths and Steyvers [9] have presented a probabilistic model of latent category topics that is similar in spirit to a semantic learning model of individual information foraging called InfoCLASS [15]. In contrast to the incremental learning process of InfoCLASS, the Topic Model [9] is an approach based on a *generative probability model*.

The Topic Model has been mainly directed at modeling the gist of words and the latent topics that occur in collections of documents. It seems especially well suited to modeling the communal production of content, such as scientific literatures [8], or wikis, as will be illustrated below.

The Topic Model assumes there is an inferred latent structure, L, that represents the gist of a set of words, g, as a probability distribution over T topics. Each topic is a distribution over words. A document is an example of a set of words. The generative process is one in which a document (a set of words) is assumed to be generated as a statistical process that selects the gist as a distribution of topics contained in the document, and then selects words from this topic distribution and the distribution of words within topics. This can be specified as

$$P(w_i \mid g) = \sum_{z_i=1}^{T} P(w_i \mid z_i) P(z_i \mid g) \qquad (1)$$

where g is the gist distribution for a document, w_i is the i^{th} word in the document, selected conditionally on the topic z_i selected for the i^{th} word conditional on the gist distribution. In essence, $P(z|g)$ reflects the prevalence of topics within a document and $P(w|z)$ reflects the importance of words within a topic.

This generative model is an instance of a class of three-level hierarchical Bayesian models known as *Latent Dirichlet Allocation* [4] or LDA. The probabilities of an LDA model can be estimated with the Gibbs sampling technique in Griffiths and Steyvers [8].

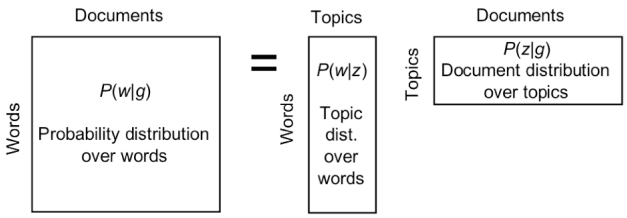

Figure 1. A spatial representation for the Topic Model

Mapping Probabilistic Semantics onto Semantic Spaces

Another pervasive way of representing semantics is to use spatial representations. Latent Semantic Analysis [LSA, 12] is one approach to semantic space representations that has had a long history of influence in human-computer interaction [3, 5-7]. LSA models words in a high-dimensional semantic space where the dimensions can be interpreted as latent topics.

Figure 1 presents a similar dimensionality-reduction formulation of the Topic Model. Topics are represented as probability distributions over words, and document gist is represented as probability distributions over topics. Theoretical evaluations that argue for the superiority of probabilistic approaches to LSA can be found in Griffiths and Steyvers [9].

Semantic Topics in Peer-produced Content

How do the semantic representations of a community engaged in social information foraging and sensemaking evolve over time? One approach to this is to apply the Topic Model to content produced by socially mediated means. Wiki's are one example of systems built to support the communal production of content. According to Wikipedia

> A wiki is a collaborative website which can be directly edited by anyone with access to is. Ward Cunningham, developer of the first wiki, WikiWikiWeb, originally described it as "the simplest online database that could possible work". One of the best-known wikis is Wikipedia

In the current modeling effort, there were two simple aims: (a) understanding the semantic topics that underlie a community wiki, and (b) understanding the changes in production of topic knowledge over time in reaction to events in the world. The Lostpedia wiki (www.lostpedia.com) turned out to be useful for this purpose. Lostpedia is a wiki devoted to users interested in the television show *Lost*, and because of the nature of the show and its multimedia channels, it is a wiki devoted to collecting intelligence from a variety of sources, and making sense of that information.

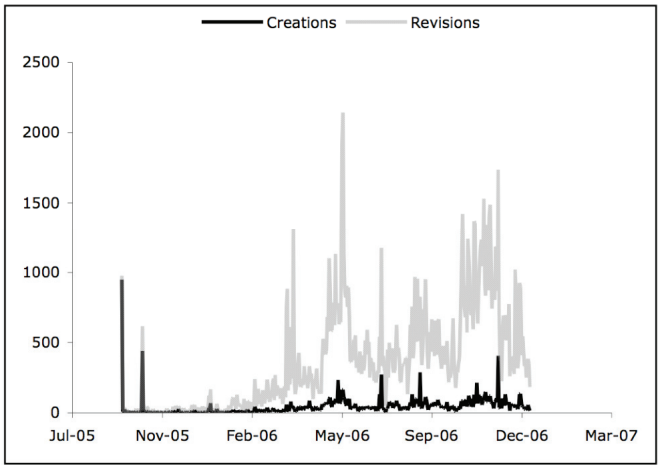

Figure 2. Daily page creation and revision rates in Lostpedia.

Lostpedia

Lostpedia was launched on September 22, 2005 at the beginning of Season 2 of American Broadcasting Corporation program *Lost*. As of August 14, 2007 Lostpedia had 3,250 articles, about 19,000 registered users, and over 92 million page views. Parts of Lostpedia are in German, Dutch, Portuguese, Polish, Spanish, French, and Italian, in addition to English. Lostpedia has been cited as "the best example of how an online community can complement a series" [11].

One reason why Lostpedia is an interesting complement to the series is the nature of the program. *Lost* is a dramatic mystery with elements of science fiction. The series follows the lives of survivors of a plane crash on a mysterious tropical island. A typical episode involves events in real time on the island, as well as flashbacks about events in the lives of characters. The fictional universe presented in the *Lost* episodes is complemented by book tie-ins, games, and an alternate reality game called the *Lost Experience*.

Crucial to the series is that it is purposely cryptic, presenting numerous unresolved mysteries and clues to their answer. This inspires a great deal of analysis and sharing of ideas and theories about the resolution of these mysteries. For instance, mysteries involve numerology (the recurring numbers 4-8-15-16-23-42), "crosses" in which characters unknowingly cross paths in their past with characters who are now involved in events on the island, and glyphs and maps which occur in many scenes. Fan theories are so central to Lostpedia that contributors must follow a theory policy. The policy discriminates between canonical and non-canonical sources. Discussion pages associated with articles are often full of uploaded frame grabs from the videos of episodes so that fans can do detailed analysis.

Lostpedia is interesting because it is very much like an open source intelligence gathering operation in which information is collected and analyzed in order to formulate

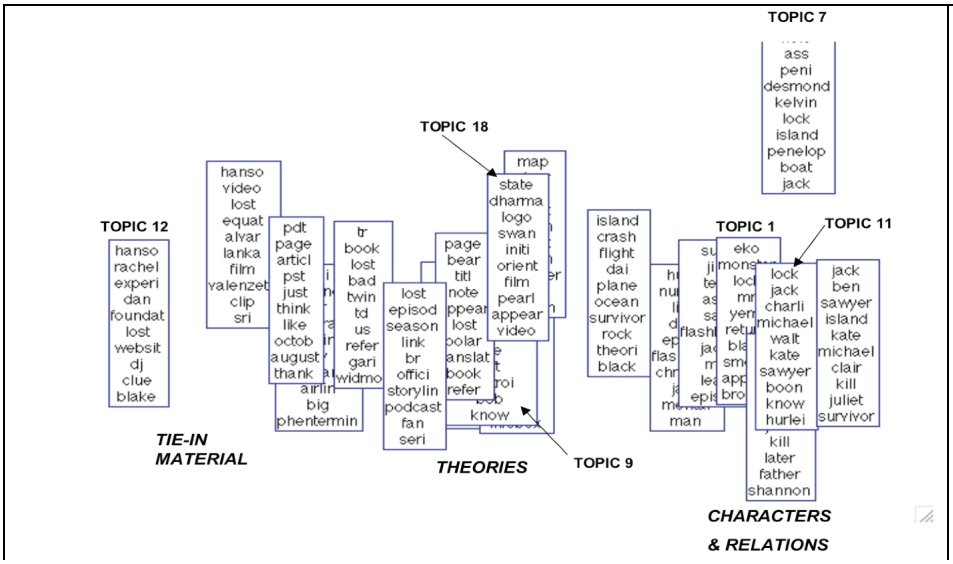

Figure 3. The Lostpedia topic space. A multidimensional scaling visualization of $T = 20$ topics in Lostpedia. Distances reflect inter-topic differences, to the extent possible in a 2D space. Topics are represented by the 10 most probable words associated with the topic. Topics in the foreground are more probable in Lostpedia.

and evaluate theories. The latent topics in this communal system surely change over time as new evidence and mysteries are presented in the world (mainly via the *Lost* program, but also via other media and information releases by writers and producers). We can therefore compare our analyses of the ebb and flow of topics in Lostpedia with what we know about events in the "real world" of releases of information in episodes of *Lost* or other media sources.

The Lostpedia Database

A Topic Model was estimated from a Lostpedia database containing page editing information. The database covered the period from September 21, 2005 to December 13, 2006. That period covers Season 2 and part of Season 3 of *Lost*.

Over that period, 18,238 pages were created and 160,204 page revisions were performed. The mean number of revisions per page was 8.78 revisions/page and the median was 2 revision/page.

Figure 2 shows the rate of page creations and revisions over the time period covered in the database. Each point in the chart is a frequency count over a day. There is a large bust of activity at the beginning of the period. Since Lostpedia was created at the beginning of Season 2 of *Lost*, there was a backlog of material to be written up. There is a large burst of activity around the Season 2 finale on May 24, 2005. There is another burst of activity following November 6, 2006, when Lost went on a mid-season hiatus until the following February (outside the range of our database).

A Topic Model for Lostpedia

The Topic Model technique used to model topics in a scientific literature [8] was used as a guide for a model of the Lostpedia database. To develop this model, the database was analyzed to produce a *word activity table*. This table was computed from all page revision text recorded in the database, passed through a stemmer and with stop words filtered out. Every Lostpedia page identified by a page identifier in the database was considered a document, and for each document, and for every day in the database, the word activity for that page was computed. The word activity for a page on a date is represented as a frequency of occurrence of words in revisions for that page on that date. The word activity table contains $D = 18{,}238$ pages and $W = 34{,}352$ distinct words.

Lostpedia Topic Space

The first goal was to represent a set of topics to encompass the span of time covering the whole available database. A word-document co-occurrence matrix was constructed by pooling all the word activity over all days. The resulting word-document co-occurrence matrix of 18,238 pages and 34,352 words contained 809,444 entries.

LDA was performed using Gibbs sampling (see the Appendix) to produce $T = 20$ topics. Exploration of other values of T (both larger and smaller) were generally consistent with the $T = 20$ model, but this number of topics is about the right size to discuss in a paper. The sampling algorithm was run for 500 iterations with parameters $\alpha = T/50$ and $\beta = 200/W$ (see Appendix).

In the results of this estimation, each topic is represented as an estimated probability distribution over words. One way to understand each topic is to examine the most probable words associated with each topic. In Figure 3, the Lostpedia topics are presented as boxes containing the 10 most probable words associated with each topic. Since the topics are represented as probabilities, the inter-topic similarities (or dissimilarities) can be defined using a KL divergence measure (see Appendix) and submitted to a multidimensional scaling (MDS) analysis that can be used to plot the topics as points in a lower-dimensional space

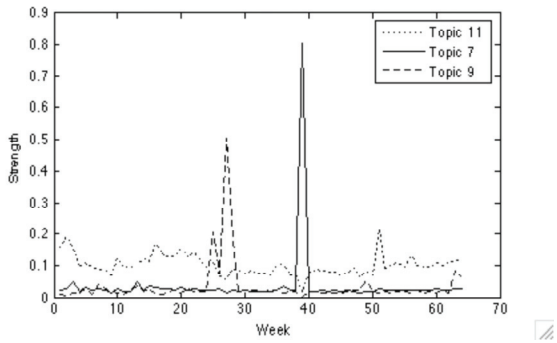

Figure 4. Stable and bursty interest in Lostpedia topics over time. Topics from Figure 3.

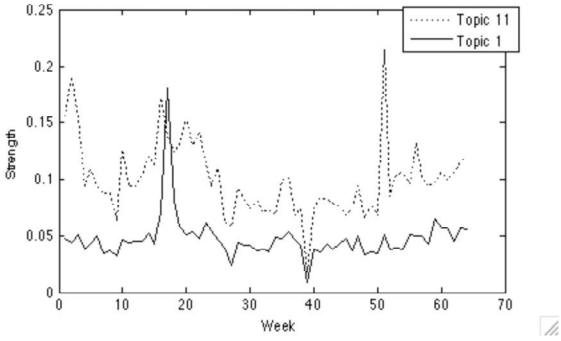

Figure 5. Appearance of a new character topic (Topic 1) causes a burst of content production.

such that the spatial distance among points reflects (as best possible) the dissimilarities among topics. Figure 3 is a plot of the 2-D MDS solution for the inter-topic similarities. The plot provides a way of visualizing a multidimensional semantic space interpretation of the Topic Model.

Interpretation of the Lostpedia Topic Space

Several regions in Figure 3 can be given a semantic interpretation as indicated by annotations to the figure. On the lower right, there is a set of topics that capture characters and relations among characters. As noted above one driving force in *Lost* is the characters, relations, crossings among characters, and flashbacks. For instance, in the center of the Characters cluster in Figure 3 is Topic 11, whose most probable words refer to the characters Locke, Jack, Charlie, Michael, Walt, Kate, Sawyer, and Boone, who are main characters, a subset of whom are referred to as the "A team", with Locke and Jack as sometimes antagonists, a Jack-Kate-Sawyer love triangle, and a Michael-Walt father-son reconciliation storyline.

In the center of Figure 3, are theories about the mysteries presented by *Lost*. For instance, Topic 18 is represented by words that refer to a fictional scientific group and experiment conducted in the past on the *Lost* island called the "Dharma Initiative", that has left behind a number of hidden stations ("Swan" and "Pearl" being the names of two stations discovered by mid-Season 3), each of which is associated with a logo, and sometimes containing orientation videos that provide clues about the workings of the Dharma Initiative and the island.

At the lower left of Figure 3 are topics that appear to be mainly associated with media tie-ins that provide information that is basically independent of the show (and especially independent of the characters) but still contribute evidence to the theories. For instance, Topic 12 is strongly associated with words representing "Rachel Blake" a fictional reporter investigating the Hanso Corporation as part of the alternate reality game called the *Lost Experience*. Although neither Rachel Blake nor the Lost Experience have much to do with the *Lost* show, the Hanso Corporation is a mysterious organization behind the Dharma Initiative and involved in the lives of several *Lost* characters.

An interesting aspect of the MDS visualization is that theories, which in the *Lost* community are a driving force, are central in the Topic Model semantic space, with topics associated with "facts" presented in the *Lost* program on one side of the theory space, and topics associated with "facts" from the *Lost* tie-ins on the other side.

Lostpedia Topic Trends

One way to assess the value of the Topic Model is to see what kinds of trends it detects in attention to topics over time, and to see how the model assessments relate to what we know about releases of information in the world of *Lost*. If there seems to be a good correspondence between what the model infers, and what we know about the progression of events in *Lost*, and assuming that people write about what is happening in *Lost*, we might gain confidence in applying this technique to other social information systems to understand the topics that people are writing about.

To analyze Lostpedia topic trends, the Lostpedia word activity database was queried to produce a set of words and word frequencies for each week in the database, with each week starting on a Wednesday and ending on a Tuesday, since episodes of *Lost* air on Wednesday nights. Each word occurrence in a weekly word profile was assigned to topics according to the probability with which it would be drawn from that topic, and for each topic the mass of word activity assigned to it was summed up. The strength of a topic was determined as the proportion of weekly word activity assigned to that topic compared to all topics.

Figure 4 presents a plot of topic strengths for a relatively stable topic over time against two bursty topics. Topic 11, as discussed above is associated with the main characters and relations. It remains very stable over the 64 week period. The topic with the biggest burst of interest among all topics is Topic 7. Topic 7 is highly associated with the words representing the characters Desmond, Kelvin, Locke, Penelope, Jack, and a boat and it is also most highly associated with Lostpedia Web pages for the Season 2 final episode called "Live together or die alone" and "Desmond

Howard", about whom it is revealed is caught in time loop. Penelope is his girlfriend, Kelvin is someone who deserted him, and Desmond was sailing a "boat" around the world to impress the father of his girlfriend, and Desmond has a great deal of interaction with Locke in the Season 2 finale. The other bursty topic is Topic 9, which is highly associated with two Lostpedia pages that apparently were discovered to contain information created by hoaxers and received substantial revision.

Figure 5 presents another bursty topic, Topic 1 is plotted against the fairly stable character Topic 11. In this case Topic 1 is primarily associated with the character Eko, who is introduced in a *Lost* episode that occurs in Week 17 of the database and remains a central character for the remainder of the database period.

The Topic Model has been applied to the analysis of scientific literatures and used to make sense of trends in the scientific literature. It appears that the Topic Model can be used to analyze how a socially mediated sense making site can be similarly analyzed to reveal the waxing and waning of interest in semantic topics in reaction to events in the world.

General Discussion

The Topic Model was applied to the Lostpedia wiki and found to reveal a sensible set of semantic topics and tracked interest in those topics in reaction to events in *Lost*. The Lostpedia wiki was a useful social sense making system to study because releases of new information in the world (from "canonical" sources including the program itself, and sanctioned media tie-ins and press) is easily determined, and the effects of these releases on the social sensemaking system can be investigated.

There are a variety of possible practical uses for models of the evolution of semantic topics in a social sensemaking system. One possible use is to provide a framework for measuring the impact of perturbations in systems on the structure and dynamics of topics, whether through new interface techniques, or policies, or community structure. Similarly such models could provide a framework for understanding how the semantic structure and dynamics have consequences for the quality of decision making , problem solving, and action.

Appendix

The Topic Model applied to the Lostpedia data used the Matlab Topic Modeling Toolbox 1.3.1 available at http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm.

Assume there are D documents, expressing T topics, and containing W words. The dataset can be represented as words

$$\mathbf{w} = \{w_1, w_2, ..., w_n\}$$

where each word w_i belongs to some document d_i. One way to represent this is as a Word-Document co-occurrence matrix containing frequencies.. An observed Word-Document matrix is viewed as the result of stochastic draws using the probabilities depicted in Figure 1

$P(w|z)$ is represented as a set ϕ of T multinomial distributions over the W words such that

$$P(w_i \mid z_i = j) = \phi_w^{(j)} \quad \text{(A.1)}$$

where z_i is the topic associated with the word w_i in document d_i and the j^{th} topic is represented as a multinomial distribution with parameters $\phi^{(j)}$.

$P(z)$ is represented as a set θ of D multinomial distributions over the T topics, such that for each document there is a multinomial distribution with parameter $\theta^{(di)}$, such that for every word w_i in document d_i the topic z_i of the word is

$$P(z_i = j) = \theta_j^{d_i} \quad \text{(A.2)}$$

The Dirichlet distribution is the Bayesian conjugate for the multinomial distribution and provides Dirichlet priors on the parameters $\phi^{(j)}$ and $\theta^{(di)}$.

Blei et al. [4] give an algorithm for finding estimates of $\phi^{(j)}$ and hyperparameters for $\theta^{(di)}$. The approach used in this paper comes from Griffiths and Steyvers [8] that employs Gibbs sampling and the use of single-valued hyperparameters α and β to specify the nature of the Dirichlet priors $\phi^{(j)}$ and $\theta^{(di)}$.

The Topic Model provides estimates of T multinomial distribution parameters, one for each topic such that the probability of any word w conditional on a topic j is defined as

$$P(w \mid z = j) = \hat{\phi}_w^{(j)} \quad \text{(A.3)}$$

The Topic Model also estimates a set of document probability distributions for over topics for each of the document d.

$$P(z = j) = \hat{\theta}_j^{(d)} \quad \text{(A.4)}$$

Acknowledgments

Portions of this paper have been submitted to CHI 2008. This research was supported by a contract from the Disruptive Technology Office. I thank Kevin Croy for providing access to the Lostpedia data, and to Bryan Pendleton and Bongwon Suh for computing the basic word activity data. Thanks also to Mark Steyvers and Thomas Griffiths for the public availability of their Matlab topic toolbox code.

References

1. Anderson, J.R. *The adaptive character of thought.* Lawrence Erlbaum Associates, Hillsdale, NJ, 1990.

2. Benkler, Y. *The wealth of networks: How social production transforms markets and freedom.* Yale University Press, New Haven, CT, 2005.

3. Blackmon, M.H., Kitajima, M. and Polson, P.G. Web interactions: Tool for accurately predicting Website

navigation problems, non-problems, problem severity, and effectiveness of repairs. *CHI 2005, ACM Conference on Human Factors in Computing Systems, CHI Letters*, *7* 1(2005). 31-40.
4. Blei, D.M., Ng, A.Y. and Jordan, M.I. Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*(2003). 993-1022.
5. Dumais, S.T. Data-driven approaches to information access. *Cognitive Science*, *27*(2003). 491-524.
6. Dumais, S.T., Furnas, G.W., Landauer, T.K., Deerwester, S. and Harshman, R., Using latent semantic analysis to improve access to textual information. in *Conference on Human Factors in Computing Systems, CHI '88*, (Washington, D.C., 1988), ACM Press, 281-285.
7. Furnas, G.W., Landauer, T.K., Gomez, L.W. and Dumais, S.T. The vocabulary problem in human-system communication. *Communcations of the ACM*, *30*(1987). 964-971.
8. Griffiths, T.L. and Steyvers, M. Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*(2004). 5228-5235.
9. Griffiths, T.L., Steyvers, M. and Tenenbaum, J.B. Topics in semantic representation. *Psychological Review*, *114* 2(2007). 211-244.
10. Jensen, M. A brief history of Weblogs *Columbia Journalism Review*, 2003.
11. Kushner, D. The Webs best, most obsessive sites for television's most additictive shows. *Rolling Stone*(2007). 34.
12. Landauer, T.K. and Dumais, S.T. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(1997). 211-240.
13. Pirolli, P. Cognitive models of human-information interaction. in Durso, F.T. ed. *Handbook of applied cognition (2nd ed.)*, John Wiley & Sons, West Sussex, England, 2007.
14. Pirolli, P. *Information foraging: A theory of adaptive interaction with information*. Oxford University Press, New York, 2007.
15. Pirolli, P. The InfoCLASS model: Conceptual richness and inter-person conceptual consensus about information collections. *Cognitive Studies: Bulletin of the Japanese Cognitive Science Society*, *11* 3(2004). 197-213.
16. Rainie, L. 28% of online Americans have used the Internet to tag content, Pew Internet & American Life Project, 2007.
17. Steyvers, M., Griffiths, T.L. and Dennis, S. Probabilistic inference in human semantic memory *TRENDS in Cognitive Science*, 2006.
18. Sunstein, C.S. *Infotopia*. Oxford University Press, New York, 2006.
19. Surowiecki, J. *The wisdom of the crowds*. Random House, New York, 2004.
20. Voss, J., Measuring Wikipedia. in *International Society for Scientometrics and Informetrics, ISSI 2005*, (Stockholm, 2005), ISSI.