# Consensus and opinions; quality and churn

**Fredrik Olsson** and **Jussi Karlgren** and **Preben Hansen**

**Martin Svensson** and **Rickard Cöster** and **Magnus Sahlgren**
SICS, Box 1263, SE-16429 Kista, Sweden
{fredriko|jussi|preben|martins|rick|mange}@sics.se

## Introduction

The role of the web user is under transformation from merely being an information consumer to also being a content provider, “from information age to participation age”, in the words of Sun CEO Scott McNealy. This increase in participation is most obviously manifested by the growth of online communities, weblogs (blogs), and various forms of cooperative and participatory publication of information.

One main factor in the shift towards participation is the advent of authoring tools for wikipedias and blogs. Such tools have decreased the threshold for publishing material online considerably — it is no longer necessary to have knowledge about the technical workings of the web to be able to use it for making information available to a massive number of potential readers. (Although the lion’s share of information produced will probably remain in text form in the foreseeable future, it should be noted that other modalities, such as podcasts, screencasts, films and images, are increasingly attracting interest.)

The dynamic nature of blogs and wikipedias poses new challenges to the field of information access and refinement; new theories, methods, and tools for alleviating the burden of digesting information on behalf of the readers are clearly needed. This paper presents some issues on readership and participation we are currently considering.

## Quality, consensus, and current affairs

A wikipedia is a highly dynamic collection of (co-) authored articles ordered by subject. Wikipedia articles may, at any time, be revised by (almost) anyone so as to reflect that persons view on the subject matter. In this sense, a wikipedia represents the consensus knowledge of a number of people on a given subject. This is reflected in the formulation and presentation, which typically assumes an encyclopaedic guise and an authoritative manner. Wikipedias purport to carry high-quality of persistent value. Reading such text, a reader should keep in mind a range of issues not necessarily as pertinent when reading traditional printed sources.

What current standpoints are there regarding the topic of a wikipedia article? Is the topic of a given wikipedia article controversial? Do the authors of a given wikipedia article have an agenda they wish to further by the text?

## Timeliness, opinions, and intellectual context

Blogs, as opposed to wikipedias, often contain highly opinionated material with strong temporal aspects; what is expressed as someone’s opinion today, may not reflect their opinion tomorrow — and may bear relatively little relation to the state of the world outside that corner of the blogosphere. A blog reader thus face a range of issues related to those a wikipedia reader faces.

What is the credibility of a given blog (post)? What other views are there regarding the subject matter of this blog (post)? What sort of social, intellectual, and factual context can this particular blog post be placed in?

## Processing models

Although we intend to address the above issues in concert, this present expose of our current work focuses on the perspective of the wikipedia reader: on how to bring more life and timeliness to the reading experience of an encyclopaedia.

Our basic idea is to enrich individual wikipedia articles with what is currently being published about the subject matter in the blogosphere. This, we believe, would provide a good ground for empowering a reader to find out what standpoints and controversies there are regarding a certain wikipedia article. The outline of the proposed method is as follows:

1. Analyse the wiki article, extract plausible keywords.
2. Expand the set of keywords to include semantically related words (synonyms, antonyms, etc).
3. Combine the keywords from step 1 and the expanded set from step 2 and find out which ones of the keywords have been used as so called (folksonomy) tags for tagging blog posts. For this, we intend to use the open API of technorati.com[1].
4. Use the valid tags from step 3 to obtain a set of related blog posts using technorati.com
5. Use the result from step 4 to mark the wikipedia article (used in step 1) with links to the most relevant blogs.

[1] Information about the API is available at http://technorati.com/developers/

Most of the technology needed to achieve the above steps is readily available, e.g., Random Indexing for finding semantically related words [2], however the combination of this technology has not yet been cobbled together into a functional architecture at the time of writing[3]. There are some building blocks that need to be tuned to this specific task – one determining reason for beginning with the wikipedia perspective is that many of them are not built to the oftentimes nonstandard writing practices of blogs.

## Example cases

To understand how wikipedia and blogs can be mixed we provide a set of cases that illustrate how wikipedia articles can be enhanced with blog content/information. These cases will serve as starting points for our evaluation.

**Extracting posts from blogs** about events dealt with in wikinews stories. Wikipedia does not only contain persistent information but also features a news site (wikinews) that is very dynamic. By extracting news headers and do a search on technorati.com it is possible to find blogs posts dealing with the same issues as the news items in wikinews — related blog posts can provide the wikinews reader with alternative views on the news item under scrutinization. In a recent wikipedia news item[4] one could read that a mutated form of the bird flu virus has been found in Turkey. A technorati search on that subject give blog entries such as whether the fears are exaggerated or the money involved in the virus[5].

**Bootstrapping wikipedia content** and providing plausible categories. Some of the featured articles in wikipedia are short, uninformative and lack appropriate categorization. As a starting point, blogs offers a simple way of providing short and uncategorized wikipedia articles with both content and categorization (matching with tags extracted from the folksonomy tags used by blog authors to categorize their posts). For example, by automatically linking to blogs about some specific subject a wikipedia article can be bootstrapped with initial content and future authors can be provided with reasonable categories.

**Retrieving information about wikipedia article authors.** An hypothesis is that many wikipedia authors are also bloggers. For many controversial wikipedia articles meta-discussions are taking place among the authors. By adding the ability for a reader to quickly find any blog posts authored or commented on by a wikipedia article author, the reader will be better equipped to judge the author's standpoint on the subject matter treated in the article. As an example consider the discussion on Opus Dei[6]. It is clear from the discussion that people have very different standpoints and to better judge the credibility of each individual author (or debater) it can be useful to find other personal views that he or she has expressed elsewhere.

## Evaluation: beyond relevance

Evaluating non-standard information access systems is a challenge in itself. Firstly, most standard metrics presuppose a statistically valid approximation to total overview of the entire collection under analysis. Such an approach is ruled out both from practical and theoretical standpoints in our case.

More importantly, the target notion of "relevance" is less practical for readership where the task at hand may not be problem solving or information gathering in the prototypical sense. "Relevance" does not take user satisfaction, quality and timeliness of texts, or reliability of authors into account; it is binary, where the intuitive and everyday understanding of relevance quite naturally is a gliding judgment; it does not take novelty, information saturation, or sequence of presentation into account.

Trying to extend the scope of an information retrieval system so that it is more task-effective, more personalized, or more enjoyable will practically always carry an attendant cost in terms of lowered formal precision and recall as measured by relevance judgments. The underlying hypothesis of our research activities is that if the concept of relevance is decomposed into its various constituent characteristics, in the present case specifically as related to the perceived sense of utility and quality on the part of the reader, we will be able to continue formal evaluation even in cases where the material at hand is dynamic, various, and fluid.

Evaluation criteria we expect to find useful, quantifiable, reliable, and valid in the present context can be packaged in an operationalization of reader-perceived *pertinence*, e.g. as measured by temporal currency of the information object as measured by its revision history by comparison to other sources; its network linkage to other information objects as determined by collection network analysis; the social characteristics of the author, as determined by social network analysis and measures of authority; textual or other intrinsic qualities of the information object, measured by stylostatistics; its similarity content- or style-wise to other objects, measured by text categorization metrics; and its similarity ecologically, measured by usage metrics. This multi-variate space of measurable characteristics of collection, author, and information object we plan to fold into a coarse estimate of reader-centered perceptual pertinence. Calibrating this measure will entail repeated empirical testing; the main thrust of our work at this point, however, is not tuning the measure to optimum performance but to find a useful framework for experimentation with represenation of textual use in face of change.

## Conclusion

This paper provides a brief outline of, and approach to, some of the research issues originating in the increase in online participation — all of which are focusing on empowering readers with means to incorporate new information into a whole in a fruitful manner.

[2] Pentti Kanerva; Jan Kristoferson; Anders Holst. 2000. "Random indexing of text samples for latent semantic analysis." In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*.

[3] October 2005

[4] http://en.wikinews.org/wiki/Sample_from_Turkish_patient_shows_mutated_Bird_Flu_virus

[5] http://fritchie.blogspot.com/2006/01/bird-flu-theres-money-in-that-there.html

[6] http://en.wikipedia.org/wiki/Talk:Opus_Dei