Context-Orientated News Filtering for Web 2.0 and Beyond

David Webster
Centre for Internet Computing
The University of Hull
Scarborough Campus
United Kingdom
D.E.Webster@dcs.hull.ac.uk

Weihong Huang
Faculty of Computing,
Information Systems, and
Mathematics
Kingston University
Kingston Upon Thames
United Kingdom
W.Huang@kingston.ac.uk

Darren Mundy,
Paul Warren
Centre for Internet Computing
The University of Hull
Scarborough Campus
United Kingdom
D.Mundy@hull.ac.uk,
p.j.warren@hull.ac.uk

ABSTRACT

How can we solve the problem of information overload in news syndication? This poster outlines the path from keyword-based body text matching to distance-measurable taxonomic tag matching, on to context scale and practical uses.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information filtering; H.3.3 [**Information Search and Retrieval**]: Relevance feedback; H.3.3 [**Information Search and Retrieval**]: Clustering

General Terms

Algorithms, Human Factors, Standardization

Keywords

Aggregation, Context, RSS, Tags, Web 2.0, Word Senses

1. INTRODUCTION

The practical use of RSS news syndication is where users can subscribe to a web site's RSS feed using a desktop news aggregator. One of the problems of desktop news aggregation is the issue of information overload[2]. If, for example, 5 RSS news feeds are subscribed to by a user, and each news feed produces 10 news items per day on average, then the user will have to filter through 50 news items in total per day. Depending on the user, this number may (or may not) be manageable. For all of RSS's convenience, the need can, be seen for intelligent filtering in client side aggregation.

2. PROBLEM SOLUTION APPROACHES

As has been previously noted, the need can, therefore, be seen for intelligent filtering on the client side of news aggregation. Existing news aggregation software lets users manually create categories and add news feeds to them. Recently, however, news aggregators, such as *NewsFire* and *NetNewsWire* have used the idea of 'smart lists', where users create categories based upon criteria and keywords they choose; these are then matched against the content of incoming news items. This is similar to setting up an email filter or virtual folder. The RSS 1.0, 2.0 and ATOM standards include category information for each news item, which enables a more elegant way to determine the topic of a news item than matching keywords against the item's text body. It is worth noting that during the time of publication of our previous paper[6], topic categorization was scarcely utilized by RSS content providers, but is now increasingly used by RSS blogging software, such as *Wordpress*.

Approaches	Level
Keyword to item body	1
Keyword to item category/topic tag	2
Namespaced keyword to namespaced item category/topic tag	3
Taxonony keyword to taxonomy item category/topic tag	4
Advanced context modelling	5

Figure 1: Approaches Stack

2.1 Word Sense Disambiguation

When a user creates a category/topic annotation for a news feed, or any other web object for that matter, word sense disambiguation becomes a concern. When a user enters the keyword into a computer, the system needs to identify and understand the sense of the word. One could easily think of three senses of the word Java for example; *programming language*, *coffee* or *island*. There is also the issue of basic level variation, polysemy and synonymy. [5] A nice quotation from Dave Winer [3], states:

You guys want users to enter metadata, I'm looking for ways to get around that, because I have found that people don't even spell things right, much less label things.

To aid in this process a public folksonomy, such as Technorati can be utilized. Similarly, a knowledge base can be used, such as the DMOZ and WordNET. Taking the coffee sense of the word Java, taking a path through the DMOZ tree would give us: *http://dmoz.org/.../Coffee_and_Tea/Coffee*.

By explicitly identifying the sense of a word, the system does not have to determine the sense of the user's category annotation or query. A problem that still remains is that of being able to communicate this to the user. It is unlikely that a user will want to search for, and enter, the path manually for every noun in their query. In order to solve this issue we have developed a graphical interface between the user and the DMOZ knowledge base.

WWW 2006, May 23–26, 2006, Edinburgh, Scotland.
ACM 1-59593-332-9/06/0005.

2.2 Context

The Oxford English Dictionary defines *context* in two ways: *the circumstances that form the setting for an event, statement, or idea, and in terms of which it can be fully understood.* and; *the parts that immediately precede and follow a word or passage and clarify its meaning.* Therefore, in the area of SW information we can say that *context* is used, not just in word sense disambiguation, but also to expand data to give it more meaning. This notion is not that different from providing *metadata* about information in an environment in order to enrich its meaning. We can therefore make the connection between context and metadata and imply that metadata can be used to provide additional contextual information for an entity.

Often related to information searching in *information retrieval*, context can be used to increase the quality of search results[8]. Recent works in this area have revolved around the utilization of context from work tasks for information retrieval and, historic contexts of the user's IR[7, 10] and acquiring context from implicit sources of evidence[1].

3. PRACTICAL CONSIDERATIONS

A simple use for RSS news topic extraction is the auto-categorization of news feeds (and also that of individual news items) in a desktop news aggregator. In this scenario, the news aggregation software requests the RSS news feed from the provider's web server. The RSS feed is then parsed by the news aggregator, which then extracts category tags from the news items. From this information, new category folders can be created in the client software and new news items can be assigned to these categories. Whilst this approach works well for simple keywords, the keyword matching system does not match contextually-related topic keywords. A combination of the idea of 'smart lists' (or a user topic interests profile) with a taxonomy-based related keyword finder the filter will provide two things: avoid polysemy mis-matches; and to match against related topics in a user profile. With the combination of taxonomy-based topic matching and a word distance measure, it is possible to measure the distance between sense-derived keywords in the user's profile and words matched in the news feed[9]. Work involving this method of taxonomy-based word distance calculation will form a significant component of a future paper we are preparing. In this approach, we will consider the discussion of context information in news aggregation from two perspectives; the client side and the service provision side. Just as in the previous example, filtering happens on the client side; this filtering could easily be transfered to a server-based solution. Just as a user can create a set of category/smart folders in his client, these options could be saved as a *user profile* and uploaded to a server. The server-based approach has the advantage of being usable with a dumb client, which can be used to subscribe to a single aggregated and filtered RSS feed. [More details on the wall poster]

4. SUMMARY AND LIMITATIONS OF THE APPROACH

The above technique illustrates one possible method of adding contextual information to an XML information source: a news feed, in the form of topic information for each news item. As can be demonstrated in the narrow domain of news syndication, there are a number of discrepant formats, each with their own methods of storing topic/category information for news items, and the feeds themselves. Our approach could be extended from simple RSS news to WWW page annotations and/or RDF document or multimedia descriptions. In order to be able to intelligently and pragmatically store and process topic/context information, we must consider two things: firstly, that topic is just one form of context information; and secondly, how to capture, then store and process context information for a web object, (such as an RSS news item) in a standardized and communicable way. For the scope of this poster, we consider context as an array of topic information for a given web object, but are actively working towards dealing with context in a much richer and high-level manner, to be presented in future work. An important direction for the approach is in testing and validating against current information filtering and text matching techniques used in RSS and email filtering. Issues that need to be considered include the quality and accuracy of the filtered result news items. A limitation of testing would be the availability of test source data, although, as previously stated, RSS providers are now starting to use topic information in their feeds. The next step will be to use namespaced tags, such as Technorati in their news feeds, with the progression to using tags from a taxonomy knowledge base such as DMOZ, which will give the advanced ability to measure the distance between tags. The scope of the approach presented here should be explored with related projects such as user interest clustering[4] and profiling to provide a piece of a much larger solution to information overload.

5. REFERENCES

[1] N. J. Belkin, G. Muresan, and X. M. Zhang. Using user's context for ir personalization. In *SIGIR 2004 Information Retrieval in Context*, pages 23–25. ACM, 2004.

[2] H. Berghel. Cyberspace 2000: dealing with information overload. *Comm. ACM*, 40(2):19–24, 1997.

[3] M. Canter. Attaching meta-data or not, 2006.

[4] D. Godoy and A. Amandi. User profiling for web page filtering. *IEEE Internet Computing*, 9(4):56–64, 2005.

[5] S. Golder and B. A. Huberman. The structure of collaborative tagging systems, 2005.

[6] W. Huang and D. Webster. Enabling context-aware agents to understand semantic resources on the wwwand the semantic web. In *Web Intelligence*, pages 138–144. IEEE Computer Society, 2004.

[7] P. Ingwersen and K. Jarvelin. Information retrieval in contexts. In *SIGIR 2004 Information Retrieval in Context*, pages 6–9. ACM, 2004.

[8] G. J. F. Jones and P. J. Brown. The role of context in information retrieval. In *SIGIR 2004 Information Retrieval in Context*, pages 20–22. ACM, 2004.

[9] S. Patwardhan, S. Banerjee, and T. Pedersen. Using measures of semantic relatedness for word sense disambiguation. In *Fourth International Conference on Intelligent Text Processing and Computational Linguistics (CICLING-03)*, 2003.

[10] I. Ruthven. and this set of words represents the user's context... In *SIGIR 2004 Information Retrieval in Context*, page 10. ACM, 2004.