Richard Freeman, "Web Feed Clustering and Tagging Aggregator Using Topological Tree-Based Self-Organizing Maps" in Proceedings of the Tenth International Conference on Intelligent Data Engineering and Automated Learning (IDEAL'09), Lecture Notes in Computer Science (LNCS 5788), Springer, 23-26 September, 2009, pp. 368-375

Web Feed Clustering and Tagging Aggregator using Topological Tree-based Self-organizing Maps

Richard T. Freeman

Capgemini, Financial Services GBU,
Technology Consulting Group,
No. 1 Forge End, Woking, Surrey, GU21 6DB
United Kingdom
richard.freeman@capgemini.com
http://www.rfreeman.net

Abstract.
With the rapid and dramatic increase in web feeds published by different publishers, providers or websites via Really Simple Syndication (RSS) and Atom, users cannot be expected to scan, select and consume all the content manually. This is leading to an information overload for consumers as the amount of content increases. With this growth there is a need to make the content more accessible and allow it to be efficiently searched and explored. This can be partially achieved by structuring and organising the content dynamically into topics or categories. Typical approaches make use of categorisation or clustering, however these approaches have a number of limitations such as the inability to represent the connections between topics and being heavy dependent on fixed parameters.

In this paper we apply the topological tree method, to dynamically identify categories, on financial and business news feed dataset. The topological tree method is used to automatically organise an aggregation of the financial news feeds into self-discovered topics and allows a drill down into sub-topics. The news feeds, organised using the topological tree method, are discussed against those of typical web aggregators. A discussion is made on the criterions of representing news feeds, and the advantages of presenting underlying topics and providing a clear view of the connections between news topics. The topological tree has been found to be a superior representation, and well suited for organising financial news content and could be applied to categorise and filter news more efficiently for market abuse detection.

Keywords: Feed aggregator, document clustering, RSS, Atom, blog, news feed, weblogs, web clustering, self-organizing maps, topological tree, neural networks, post retrieval clustering, taxonomy generation, enterprise content management, enterprise search, information management, topic maps.

1 Introduction

The rapidly growing volume of dynamic content on the Internet is leading to an information overload, where users have to be selective on what they choose to read. They need to quickly scan and select interesting topics in the title and avoid duplicate articles from different publishers as they have already consumed this information. The

amount of content has also been augmented, over the past few years, with the increased popularity of user generated content such as weblogs / blogs (e.g. TypePad[1]), micro-blogs (Twitter[2]), photo sharing services (e.g. Picasa[3]) and social networking (e.g. Facebook[4]), as well as prevalence of web feeds from various websites. For example blog search engine Technorati[5] reports about 1 million posts a day and more than 200 million blogs in the world today, and LiveJournal[6] reports about 20 millions accounts. This user-generated content is rapidly expanding as authors can touch a large number of readers with at low cost, and with little publication and marketing effort. In addition, another trend is the growing number of users that are attracted by content, rather than regularly returning to visit particular websites, i.e. they navigate directly to the article of interest either via a search engine or a web feed reader.

Web Feeds permit the subscription to regular or frequently updated content. Web content can be pushed to the subscribers, where they can subscribe to any type of web feeds of interest to them and can unsubscribe at any time. For example a user would not typically navigate daily to a blog, but would instead subscribe to be notified by email of any additions by RSS. Web feeds are not limited to news content, they include anything that is periodically updated or with a notification such as podcasts, blogs, social networking, social photo-sharing service, or publication updates. Web feeds allows users to subscribe to syndicated headline or headline-and-short-summary content which can be delivered to different channels including browsers, web portals, news readers, email, mash-ups, widgets / gadgets and mobile devices. A link to the full content as well as other metadata is also generally provided in the feed. This format has the advantage to also be machine readable allowing it to be for example embedded in other websites or collected by feed aggregators.

Figure 1 – Example web feed publishers and readers

Web feed aggregator applications can pull relevant content to which the user has subscribed to. This allows the web feeds to be read from within one location or

[1] http://www.typepad.com/
[2] http://twitter.com/
[3] http://picasa.google.com/
[4] http://www.facebook.com/
[5] http://technorati.com/
[6] http://www.livejournal.com/

program. Existing schemas such as RSS do provide limited metadata (e.g. date, title, and summary).

News feed aggregators include a combination of web feeds and often screen scrappers which index news websites directly. To organising the content into different categories is generally a manual task (making them potentially inconsistent across different web feed sources). For example news aggregator websites such as the Drudge Report[7] and the Huffington Post[8] are manually organised and moderated by the website owners. However websites like Google News[9] or Techmeme[10], aggregate new content autonomously using algorithms which carry out textual analysis and group similar stories together to minimise duplicates. However as we will discuss later, these categories are limited, do not allow drill down by topic and the user has little control on which web sites the online news aggregator subscribes to.

Hence there is a need to organise these news stories dynamically and more efficiently with minimum human intervention. The purpose can be to find related stories or categories which can be browsed by a human user or used by a machine to perform textual analysis.

In this paper we propose a novel approach to news aggregation using self-organising map-based topological tree method to organise the content into a human friendly hierarchical representation, which adapts to the natural underlying topic structure. RSS feeds are often focused on one area such as business or technology, and if a user subscribes to many publishers for the same topic it is often beneficial to remove duplicates and be able to dynamically drill down on the main topics of business or technology. Such hierarchies or tags would be difficult to keep up to date by human editors and might be subject to human (mis)interpretation. If this was manually done then if would require a large number of contributors to create and maintain the tags, e.g. Delicious[11] has five million users. Allowing the user to choose the feeds of interest and organise then using a topological tree allows the users to see the dominant topics and how they are related by looking at neighbouring nodes [1]. These topics are automatically discovered based on the content and metadata present in the feed and is organised so that similar topics (or nodes) are located close to one another. Since the metadata in the RSS / Atom feeds have mostly limited metadata such as date, description and title, the proposed system also crawls and indexes the actual content page.

2 Visual Representation of Web Feeds Content

2.1 The Importance of Clustering and Topology

Typical aggregators collect the feeds into folders based on the feed URL and the user is allowed to manually create some folders to group them together, e.g. technology. The main limitation is that as the number of subscriptions grows, from different

[7] http://www.drudgereport.com/
[8] http://www.huffingtonpost.com/
[9] http://news.google.com/
[10] http://www.techmeme.com/
[11] http://delicious.com/

publishers, the number of folders and content also increases. This makes it difficult to be selective on what to read, e.g. users will have to quickly skim each headline one by one to determine if it is of interest to read. In addition there will be some duplicates from different providers. To address this, some aggregators such as Newz Crawler have the concept of Smart Folders, which can be thought as populating a folder by selecting specific feed items from all publishers that match a pre-stored search query. Some approaches make use of content-based filtering (based on the patterns in the contents) or collaborative filtering (based on communities or user preferences) [2] or a mix of both [3]. Some of the content-based filtering approaches use document clustering [2][4][5], extract taxonomy or semantic tags [6] [7] or feed-based time series [8]. However the benefits of unsupervised neural networks such as the non-linear input-output mapping, and ability to learn from experience and examples make them well suited for content categorisation where the underlying topics are not known. The two major types of neural networks used for document organisation are the adaptive resonance theory (ART) [1] and self-organising maps (SOM) [9]. SOMs are generally associated with 2-dimensional structures that help visualise clusters and their relationships in a topology. However 1-dimensional SOMs can also be used to visualise data. The topological tree method uses 1-dimensional SOMs or growing chains, where each node may spawn a child chain effectively constructing a hierarchy chains allowing a parent-to-child drill down. The Topological Tree structure has already shown to be an efficient representation against existing post-retrieval clustering search engines such as Vivisimo [10] and social networking search [11].

3 The Topological Tree Method

3.1 Pre-processing

In the pre-processing stage the content in the dataset is crawled, parsed and stemmed. The remaining terms are filtered so that only the most discriminate terms are stored in a compressed vector space model representation (VSM) [12]. In the VSM a document is stored as $[f_1, f_2, ..., f_N]$, where f_i is the frequency of term i and N is the total number of unique terms. Since a document vector is sparse, its term frequencies can be stored in a compressed representation as a hash table for efficient retrieval,

$$\mathbf{x} = [\{\gamma_1, \rho_1\}, \{\gamma_2, \rho_2\}, ..., \{\gamma_r, \rho_r\}], \ \gamma_k \in \{V \mid f_{\gamma_k} \neq 0\}, \ \ k = 1, 2, ...r \quad (1)$$

where $\mathbf{x}$ is a document vector stored as term γ and weighted frequency ρ pairs, γ_k refers to the γ_k-th term in the vocabulary V, r is the total number of terms in the document. Term weighting ρ and feature selection are detailed in [13]

3.2 Growing Chains and Topological Tree Method

The Topological Tree method uses a set of chains adaptively developed by a growing chains (GC) algorithm. Each chain begins with two nodes and grows until the termination process stops it [14]. There are two important steps in the GC training: the search for the best matching unit and the update of the weights of the winner and its neighbourhood.

The best matching unit of c(**x**) is the node with maximum S_{dot} amongst all nodes n with respect to a document vector $\mathbf{x}(t)$,

$$c(\mathbf{x}) = \arg\max_j \left\{ S_{dot}(\mathbf{x}(t), \mathbf{w}_j) \right\}, \quad j = 1, 2, ...n \tag{2}$$

Once the winner node $c(\mathbf{x})$ is found the neighbouring weights are updated using,

$$\mathbf{w}_j(t+1) = \frac{\mathbf{w}_j(t) + \alpha(t) h_{j,c(x)}(t) \mathbf{x}(t)}{\left\| \mathbf{w}_j(t) + \alpha(t) h_{j,c(x)}(t) \mathbf{x}(t) \right\|} \tag{3}$$

where $\alpha(t)$ is the monotonically decreasing learning rate and $h_{j,c(x)}(t)$ the neighbourhood function, typically a Gaussian kernel. When the learning has stabilised for the current number of nodes *n*, the entropy of the chain is recorded and a new node is inserted next to the node with the highest number of wins. A validation criterion, termed entropy-based Bayesian Information Criterion that penalises complexity, is used to estimate the optimum number of nodes per chain [14].

In the hierarchical expansion process, each node in the chain is tested to see if it will spawn a child chain. This is performed using several threshold and cluster tendency tests. If any of these tests fail for a particular node, then it does not spawn a child chain and becomes a leaf node in the final topological tree representation.

Finally each node in the chain is labelled using the most representative terms of the node's weight and its frequency. Once the chain is labelled, then it is added to the current topological tree structure. If further hierarchical expansions in its child chains are required, then the process is repeated for each of the child chains, otherwise the process is terminated and the results presented to the user.

3.3 The Topological Tree Feeds Aggregator System

In the Topological Tree feeds aggregator system the feed content is first captured via the RSS / Atom adaptor then the full text content is retrieved via the HTML Adaptor. Both adaptors are called by the feed crawling service which has for task to run though all the candidate feeds the user has selected. The content and feeds are then analysed by the pre-processing services as described in 3.1 and clustered using the ATTS and post-processed as described in 3.2. The User Interface layer allows the user to configure the subscription, crawling, pre/post-processing and training services, as well as explore and search the output results. A diagram showing the overall architecture is shown in Figure 2.

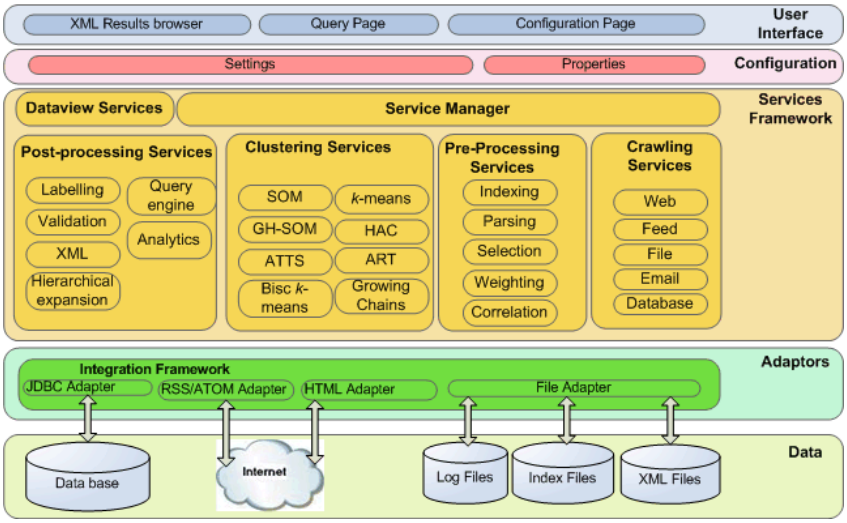

Figure 2 – Adaptive Topological Tree Feeds Aggregator Architecture

4 Results and Discussions

Newz Crawler, NewsGator, Google Reader both have no categories options but instead allow the users to create a set of folders. Google News aggregates news feeds automatically to a limited set of predefined categories of top stories, world, U.S., business, sci/tech, entertainment, sports, health, or most popular. For example some viewers might not be interested in US Sports, and the only way to filter this down to Welsh sports is to run a generic search. In addition some sites such as Yahoo Finance have feeds that are already categorised into sub-categories such as Most Popular, Bonds, Commodities, Currencies, Funds, IPOs etc. Unfortunately this metadata is not typically passed to the news aggregator.

In the experiments, the dataset was dynamically generated from RSS news feeds relating to business and finance from a recent mix of Reuters, AFP, Sky Business News, Yahoo Finance and ShareCast content publishers. The results for the topological tree are shown in Figure 4. In Google News, Google Reader and Newz Crawler the results are difficult to browse as the volumes increase, and it is difficult to identify different types of finance or business sub-topics they might be interested. In comparison, the topological tree representation appears more intuitive and natural to the user, as closely related topics are located close to one another in each chain. In addition hierarchical relations between a parent node and child chain help abstract different levels of detail. Typical aggregation approaches are compared and discussed in Table 1.

Existing market abuse solutions could also benefit from the proposed approach as it can be applied to less structured news, e.g. blogs, webfeeds, social networking sites from different vendors etc. Each of the detected categories (representing an aggregation from different publishers) could be given a suspicion rating, hype factor or other metrics that could be used for market abuse detection such as in insider dealing where we are looking to detect leaking of information, and parties who have acted on non-public or sensitive information.

For more details and enhancements please refer to the journal papers listed on http://www.rfreeman.net/

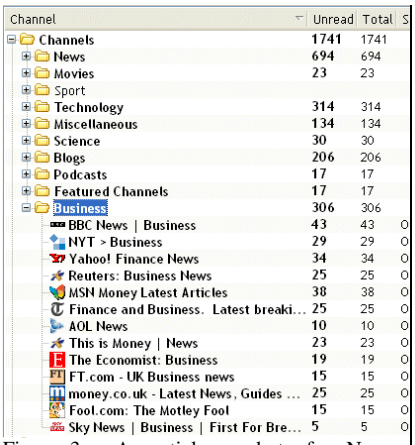

Figure 3 – A partial snapshot of a Newz Crawler tree showing the large number of newsfeeds organised by provider. Here users have to quickly evaluate if an article is of interest or not or if they have already read something similar from another publisher. As the number increases the user get submerged with feeds.

Figure 4 – A topological tree generated from the new business and finance feeds. Folders are created based on the contents of the news feeds and the topology ensures that similar topics are closely grouped together making it more intuitive to navigate. From top to bottom the following topics can be observed: automotive (GM), energy, dollar, iron ore (Rio Tinto), mergers and acquisitions, UBS, Oil Prices.

Table 1. Comparison of the differing methods for representing feeds content.

Name	Description	Issues
Aggregated inbox	All the feeds are stored in an inbox, e.g. Google Reader. Others use a ranked list	If the user subscribes to many sources the inbox will very quickly be overloaded
One feeds per folder	All the feeds are stored in it's own folder, e.g. NewsGator supports this	If a user subscribes to hundreds of feeds then they will have to view each folders contents individually which is not ideal.
News aggregation by topic	Website that offer aggregated news from other news websites or news feeds.	Users cannot select which feeds to subscribe to, topics are limited or assigned manually
Clustering feeds*	Feeds can be clustered (currently not typical in aggregators), i.e. grouped together by similarity.	Dependent on parameters (e.g. number of clusters) and cannot represent connections between topics
Community driven tagging*	Supporting community recommendations or views (currently not typical in aggregators)	Depends on the community supporting the tagging the articles / feeds and the reader / aggregator supporting the metadata
Virtual folder / personalised view*	Virtual folders can be created which only select articles meeting some search criterion, e.g. Newz Crawler supports this. Alternatively the system could be trained with articles of interest (currently not typical in aggregators) in a similar way to document classification.	Limited flexibility, does not dynamically adapt to latest topics
Proposed Adaptive Topological Tree feed aggregator	The user chooses to subscribe to the web feeds of interest and runs the system. The output results are displayed as a tree view where each node is a topic folder. In this tree similar topics are located close to one another due to the inherent self-organising topology.	Users need to initially configure some of the settings and feeds to subscribe to.

*The table above also indicates limitation of features of future feed aggregator features that are not currently typical but may become more common in the future.

5 Conclusion and Future Work

We have presented a self-organising method for generating a Topological Tree structured representation from web feeds, where the size of each level is determined through validation. Compared to the existing methods the proposed Topological Tree is adaptive, dynamically validated and is natural for visualisation and exploring as in many web directories. The topology provides a unique characteristic that can be used for finding related topics during browsing and extending the search space, a clear distinctive advantage over other feed or news feed aggregators, or clustering approaches. The topological tree approach could be useful in detecting market abuse by better aggregating feeds from many publishers and organising them. For example, the proposed approach could help assist detecting insider dealing, which is typically centred around looking at the release of sensitive news content and actions that were taken around the time of the announcement. Future releases will focus on combining the Topological Tree method with existing taxonomies-based and collaborative filtering approaches.

References

[1] R.T. Freeman: Web Document Search, Organisation and Exploration Using Self-Organising Neural Networks, *PhD Thesis*, Faculty of Engineering and Physical Sciences, School of Electrical & Electronic Engineering, University of Manchester: Manchester (2004)
[2] A. Qamra, B. Tseng, E.Y. Chang: Mining blog stories using community-based and temporal clustering. In: 15th ACM international conference on Information and knowledge management CIKM, pp – 58-67, ACM (2006)
[3] G. Paliouras, M. Alexandros, C. Ntoutsis, A. Alexopoulos, and C. Skourlas: PNS: Personalized Multi-Source News Delivery. In: 10th International Conference on Knowledge-Based & Intelligent Information & Engineering Systems. LNCS, vol. 4252, pp. 1152 – 1161. Springer (2006)
[4] X. Li, J. Yan, Z-H. Deng, L. Ji, W. Fan, B. Zhang, Z. Chen: A Novel Clustering-based RSS Aggregator, In: 16th international conference on World Wide Web, pp. 1309 – 1310. ACM (2007)
[5] N. Agarwal, M. Galan, H. Liu, S. Subramanya: Clustering Blogs with Collective Wisdom, In: 8th International Conference on Web Engineering, ICWE'08. pp. 336 – 339. IEEE (2008)
[6] W. Huang, and D. Webster: Enabling Context-Aware Agents to Understand Semantic Resources on The WWW and The SemanticWeb, International Conference on Web Intelligence (WI'04), pp. 138 – 144. IEEE (2004)
[7] D. Webster, W. Huang, D. Mundy, P. Warren, Context-Orientated News Filtering for Web 2.0 and Beyond, In: 15th International World Wide Web Conference, pp. 1001 – 1002. ACM (2006)
[8] M. Thelwall, R. Prabowo: Identifying and Characterizing Public Science-Related Fears From RSS Feeds, Journal of the American Society for Information Science and Technology, 58(3), 379-390 (2007)
[9] T. Kohonen: Self-Organizing Maps, Third Extended Edition, Springer (2001)
[10] R.T. Freeman: Topological Tree Clustering of Web Search Results. In Intelligent Data Engineering and Automated Learning, IDEAL 2006, LNCS, vol. 4224, pp. 789-797, Springer (2006)
[11] R.T. Freeman: Topological Tree Clustering of Social Network Search Results. In: Intelligent Data Engineering and Automated Learning - IDEAL 2007. LNCS, vol. 4481, pp. 760-769. Springer (2007)
[12] G. Salton: Automatic text processing - the transformation, analysis, and retrieval of information by computer. Addison-Wesley (1989)
[13] R.T. Freeman, and H. Yin: Web content management by self-organization. Neural Networks, IEEE Transactions on. **16**(5), 1256-1268 (2005)
[14] R.T. Freeman and H. Yin.: Adaptive topological tree structure for document organisation and visualisation. Neural Networks. **17**(8-9), 1255-1271 (2004)