NAVIGATING THE BLOGOSPHERE USING CONTENT CLOUDS

Robert Burns and Kenneth Cosh

Computer Information Systems,
Payap University,
Chiang Mai, Thailand.

ABSTRACT

The term 'Blogosphere' is given to the wide variety of interrelated blogs existing across the Internet. Traversing and exploring this blog space can become a disorienting and confusing experience as there is generally little navigation support available. As the era of Web 2.0 has emerged, there has been increasing amounts of user generated content, blogs are just one example of the content being produced. This paper highlights some managerial challenges concerning large, diverse amounts of user generated content, such as classification and navigation. We present 2 novel methods to aid with this management; Firstly Content Clouds, which assist with content classification and present a visualization of some content. Secondly a Navigation Tool which can be used to traverse collections of data – and we demonstrate this tool on a collection of blogs.

Index Terms— Web 2.0, Blogging, Data Visualization, Natural Language Processing

1. INTRODUCTION

The internet and related communication technologies have had far reaching effects on a wide variety of industries with the opportunities for e-Business. The world is becoming more digital, affording new forms of virtual lifestyle, created by increasing electronic activities, from e-commerce to e-learning, from e-banking to e-societies. Digital homes and digital libraries are all supported by modern information and communication technologies.

More recent developments have been dubbed Web 2.0, marking a new era of the internet. Technological and conceptual evolutions have created an architecture of participation, in which users are expected to create and add to the web's content. Traditionally webmasters produced content to be read by their visitors, while in Web 2.0, this content is often produced by the site's visitors. Content can take the form of comments on blogs, discussion forums, and there is even more interactivity through community sites and wikis, where users have the ability to contribute, change and update the collective knowledge contained within[12].

Web 2.0 is a controversial term, referring to a perceived new generation of internet-based services. Programming approaches, such as AJAX, afford web developers new capabilities allowing them to offer innovative products and services. Regardless of whether Web 2.0 is a new generation of the internet, there is a noticeable trend towards users participating more actively in the websites they visit. Popular new activities have emerged such as blogging, and blogging in turn has facilitated novel social interactions[10]. Blogs have also become an important eBusiness tool that enables organizations to interact with their customers and potential customers in a unique way.

In this paper we discuss some of the issues that have arisen due to a need to manage increasing amounts of user generated content; such as classification, navigation and control issues. We present Content Clouds a novel way to classify user generated content quickly and automatically and a means to visualize the content that has been generated. We also present a navigation tool developed using Content Clouds that can be used to navigate a collection of documents. For the purpose of demonstration a collection of blogs about travel in Thailand was used as content.

2. MANAGING WEB 2.0 CONTENT

Web 2.0 encourages an architecture of participation, where website visitors are encouraged to contribute to the website's content. As more content is created each day, some managerial challenges have arisen. This paper will address some concerns with the current means of management.

2.1. Classification

One of the challenges for dealing with increased amounts of content is classification. Traditionally content is classified through a taxonomy, which ensures that related content can be located together. Managing a taxonomy is more difficult when the content is produced minute by minute throughout the world, in multiple languages, using varied media, to different quality levels, by a huge, diverse population. One accepted way of classifying this content is to pass the responsibility for classification over to a community, along with the responsibility for creating and maintaining the content. The community then decides where the content belongs, normally through 'tagging'. Here, rather than using a taxonomy, a folksonomy is generated [9][5]. Once content has been classified by collective tagging, a variety of semantic information can be deduced from this metadata.

One application of tags is creating a tag cloud (see figure 1), a simple visualization of the most frequently used tags [Godwin]. If the content on a website is continually tagged with a certain subject, it is assumed that the site concerns the subject and the tag is presented more visibly. The tag cloud can then be used to aid searching on a particular site. A variety of studies have evaluated the creation and use of tag clouds from a usability perspective [15][6].

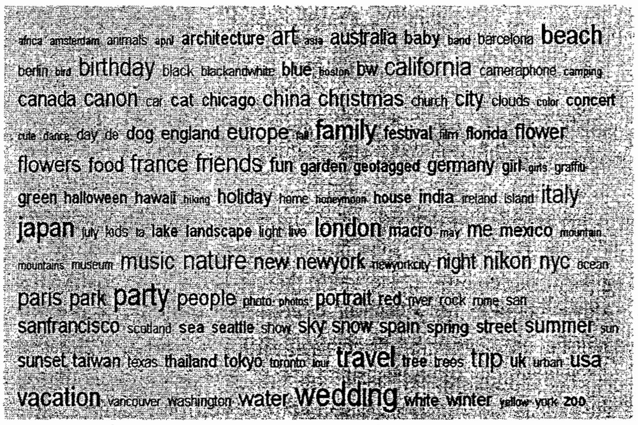

africa amsterdam animals april architecture art asia australia baby band barcelona beach
berlin bird birthday black blackandwhite blue boston bw california cameraphone camping
canada canon car cat chicago china christmas church city clouds color concert
cute dance day de dog england europe fall family festival film florida flower
flowers food france friends fun garden geotagged germany girl girls graffiti
green halloween hawaii hiking holiday home honeymoon house india ireland island italy
japan july kids la lake landscape light live london macro may me mexico mountain
mountains museum music nature new newyork newyorkcity night nikon nyc ocean
paris park party people photo photos portrait red river rock rome san
sanfrancisco scotland sea seattle show sky snow spain spring street summer sun
sunset taiwan texas thailand tokyo toronto tour travel tree trees trip uk urban usa
vacation vancouver washington water wedding white winter yellow york zoo

Fig. 1 A Tag Cloud from Flickr.com

There are a number of issues with tag cloud creation, which this paper will address. First, to effectively create a tag cloud, the community needs to have effectively tagged the content and for that a willing and capable community is needed. Secondly a collection of content is necessary as it isn't possible to create a tag cloud based on individual items of content (such as a single photo, or a single article of text). In section 3 we examine how to automatically generate an effective cloud, based on individual pieces of text without the need of a folksonomy, by applying Natural Language Processing (NLP) techniques.

2.2. Navigation

The internet is a large space of web pages each of which is complexly interrelated through hyperlinks. The web page follows the metaphor of the physical pages of a book or text, but the difference in navigation is crucial – while texts are navigated linearly, web pages are not. Links between web pages are based on words or phrases within the content of a page, with the semantic relation between the target and source being entirely dependent on the source. Because of this, users don't know where a link will lead until they click on it [11]. Finding a particular piece of information may require a protracted search, including backtracking to previous pages, and traversing multiple pages [2].

For the most part, search engines operate as a portal to the vast information space of the web, supporting the searching aspect of web navigation. Once a link has been followed users can feel lost and disoriented, especially with sites that offer little or no navigation support. With increasing amounts of user generated content, searching, filtering and traversing the content becomes increasingly more complex. While spiders can crawl and organize the content of the web relatively successfully, as that content continually changes and is updated there is no guarantee that the content will remain as expected.

2.3. Quality

The traditional model of print publication involves a qualified editor and for academic publications a rigorous peer review process. In the early days of the web, to 'publish' content on a web page, all that was needed was a domain and a host. With web 2.0 the ability to publish content has become even more widespread. Wikipedia allows any visitor to edit the content of an article, regardless of their edits accuracy. While for this site a community is in place to monitor and revert false edits, the approach risks confusing fact with popular opinion [18]. Other sites have less error checking and prevention.

While knowledgeable editors exist and attempt to improve the overall quality of web 2.0 content, others amplify mistakes through ignorance, sloppy research, malice or zeal. Some web surfers are trained to thoroughly research, analyse and evaluate their findings, others have no way of knowing which kind of article they have found, whether it is accurate or not. While the age of participation encourages casual editors to be bold and to create and publish content, there is also a need for monitoring and quality assurance.

2.4. Control

From a business perspective, there are also concerns about controlling the way an organization is presented. Companies go online, by producing a webpage about their products or services, carefully branded to project their desired image. With increasing participation, other web users can present a very different image of an organization. Customers and potential customers, rather than seeing the desired image, could stumble upon a blog entry from a dissatisfied customer or malicious competitor. Due to the challenges of navigation, and quality, highlighted in the previous sections, a company may never be aware of a negative presentation about their organization.

Recent research showed that there were significant differences between the branding of a travel destination and the perception of that destination by its customers. For example the tourist resort of Pattaya is branded by the tourism authority as a destination for water sports, while customers perceive it as a sex tourism destination[1]. Managing this difference is a further challenge presented by web 2.0.

3. CONTENT CLOUDS

The challenges presented in section 2 were our motivation for developing some tools to address at least some of the issues. Tag Clouds are used by many sites to indicate their key topics, and they offer a simple but effective visualization of that meta data, but with some limitations. First, they can only be applied to collections of data, rather than individual documents, and secondly they require a reliable and competent community of taggers. Applying some natural language processing tools can address both of these issues and generate a Content Cloud.

3.1. Natural Language Processing

NLP can use computers to extract and analyse information from a natural language source, such as an article generated by an internet user. There are 2 key approaches to NLP, rule based and probabilistic. For this technique a probabilistic method is applied, our objective being to take a piece of text and extract the key words, which would ordinarily be used as tags, along with an appropriate weighting for each word. Probabilistic NLP has been used in a variety of disciplines to gain an understanding of a document's contents automatically. The REVERE project applies NLP to extract requirements from legacy documentation [14]. As well as extracting requirements, alternative models can also be created using NLP tools. NLOS creates a semiotic model of an ethnographic report, with reduced analyst involvement [3].

3.2. Creating Content Clouds

Content Clouds are based on the content of a piece of text, such as a blog entry, therefore the approach involves applying NLP tools to the content, rather than the tags a community has assigned to the content. The first step is to identify the keywords used within the text – a document about the stock exchange is more likely to contain words such as 'stock' and 'banking' than it is to contain 'badger' and 'stoat'. A word frequency list for each article can easily be generated. The most common words in most documents are the same – words such as 'the', 'of' and 'and', but these words rarely add to the understanding of an article.

One approach for eliminating these words is to use a simple skiplist, which lists words to be ignored. This has 2 key drawbacks, first that a document describing the word 'the' would simply ignore each occurrence of the word. Secondly, marginal but frequent words that don't occur on the skiplist often emerge with a high rating. Our approach compares the article's frequency list with that of a standard word frequency corpus, to identify the statistically significantly overused words in the article. The British National Corpus (BNC) is a large collection of words used in the English language, collected from a diverse selection of sources representing a wide cross section of domains[7]. The first step to identifying the significantly overused words, is to calculate how frequently each word should be expected to occur. This can be done by using the following contingency table for each word in the document.

	Text to be analysed	BNC	Total
Frequency of word	a	b	a + b
Frequency of other words	c-a	d-b	c + d – a – b
Total	c	d	c + d

In the contingency table, 'a' represents the frequency of the word in question, and 'b' its corresponding frequency in the BNC. 'c' represents the total number of words in the document, and likewise 'd' the 100 million words included in the BNC[14]. The expected occurrence value of each word is calculated using the following formula, in which 'O' represents the 'observed' values, or with reference to the contingency table 'a + b', 'N' represents the total values or 'c + d', and 'E' represents the expected value. 'i' is 1 for the text to be analysed, and 2 for the BNC.

$$E_i = \frac{N_i \sum_i O_i}{\sum_i N_i}$$

To simplify the formula for each individual case, the expected value for any word in the document, given its frequency in the BNC is calculated from the values in the contingency table, using the following algorithm.

$$E1 = c * (a + b) / (c + d)$$
$$E2 = d * (a + b) / (c + d)$$

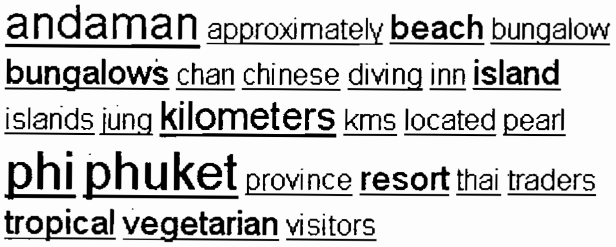
Given the expected frequency for each word, and the observed frequency for each word, the likelihood of that result having occurred can be calculated using the log-likelihood calculation, as demonstrated in the following formula.

$$-2\ln\lambda = 2\sum_i O_i \ln\left(\frac{O_i}{E_i}\right)$$

This formula can be rearranged to calculate the log-likelihood (LL) using the following calculation.

$$LL = 2*((a*\ln (a/E1)) + (b*\ln (b/E2)))$$

The higher the LL result is, the more significant the difference between the expected value and the observed value is. Using statistical tables, significance at the 5% level requires a LL value of greater than 3.8. To be significant at the 1% level the LL would be greater than 6.6. Therefore, the most significantly overused words are those recording the highest LL. From this list of overused words a cloud can be created, using the LL score to decide on the size of each word. In our tests we have chosen the 25 words with the highest log-likelihood. In order to display them in a visually meaningful way, we normalize the those 25 values over the range 1 – 1.8 at an increment of 0.2. The resulting values are used to size the display font in 'em' units. Figure 2 shows a sample content cloud created from a blog entry on the history of Phuket.



andaman approximately beach bungalow bungalows chan chinese diving inn island islands jung kilometers kms located pearl phi phuket province resort thai traders tropical vegetarian visitors

Fig. 2 – A Content Cloud about Phuket

3.3. Uses of Content Clouds

Originally Content Clouds were intended as an automatic replacement for Tag Clouds, but further evaluation has shown that Content Clouds offer an alternative, yet complimentary classification of content. We found that often the words appearing in a Content Cloud, surprisingly didn't appear in the corresponding tags. One reason for this is that currently the Content Cloud algorithm works at a lexeme level, only allowing individual words, while tagging allows word pairs or phrases – notice how "Phi" appears prominently in the content cloud in figure 2, after the island "Ko Phi Phi". More significantly, Content Clouds work at a different level of abstraction to tag clouds.

Saussure [17], when studying semiotics, noted a two way relationship between the form a sign takes (signifier) and the concept it represents (signified). There is often a difference between the signifier (in this case the text of an article), and the signified (the content). Peirce [12] added a third aspect to the relationship, that of the interpretant – the way the sign is interpreted by the receiver. The interpretant can also be different, in our case the tag which reflects the interpretation by the reader. This variation need not cause concerns in terms of effective communication.

Signs can be understood at different levels. At a syntactic level, the syntax of a sign and its relationship with other signs is considered; essentially the grammar of an article. At a semantic level, the meaning of individual elements of the sign are interpreted. Natural language affords a rich diversity of words which can be to some extent interchanged or substituted or swapped. At a pragmatic level the intended message is considered, using not only the sign, but related external knowledge, experience and culture. Pragmatic interpretation allows the sign to take a very diverse form and still be correctly interpreted.

Both the content and the associated tags are used by the author to convey a message (sign) to the reader. Tags are constrained to one or two word phrases, which forces the tag creator to be precise and unambiguous. Contrastingly, content can make use of a much broader language, so it is likely to use a different vocabulary. The purpose of both the Tag Cloud and the Content Cloud is to be an aid to classification and navigation, representing a conceptual namespace. Both visualizations offer related information so can be considered complementary classifications.

4. WEB BLOGS

Blogs have become an increasingly popular means of publishing content to the web and blogs are used for a wide variety of activities, including publishing a personal diary, and a means of finding a community of like minded individuals to associate with[16]. One of the key aspects of blogging is the simplicity of use, enabling even those with the most basic knowledge of computers to contribute, without the need to understand HTML or PHP[8].

Blogs are also used by businesses as a marketing tool, allowing internet based press releases and an alternative means of connecting with their customers and directing them towards their products and services through both online and offline channels. Similarly blogs are used by customers to review their experiences and offer opinions on the quality of goods and services. Another popular area of blogging is that of travel blogs, with tourists writing a blog of their travel experiences to easily update their friends and relatives back home. The navigation tool discussed in the following sections is demonstrated on a collection of travel blogs about Thailand.

4.1. Navigating the Blogosphere

The Blogosphere is a term used to describe all the blogs available and their interconnected nature. Due to the hyperlinks between blogs, and many blogs being hosted on the same site (such as blogger.com) there is a perception that all blogs are somehow interrelated within some connected community. This community is further reinforced through the ability to comment on blogs, and references by way of trackbacks, which are used when one blogger writes an article which builds upon another. While the blogs might exist within the same conceptual space, the blogosphere, navigation of this space is somewhat restricted. A few options are available;

First, some blogging sites allow a random jump – where clicking 'next blog' will teleport the user to a random unrelated blog somewhere else within the blogosphere. While this can be fun for those with time and no fixed objective, it can prove to be disorienting and certainly isn't an effective means of navigating to specific content. The random jumps are also limited to the site in which the currently viewed blog is hosted. Many blogs also contain a 'blogroll' set of links, which are links to related blogs – written by other members of the community. As it is chosen by the blogs author, it is likely that the linked to content will be somehow related, but again it isn't a sophisticated navigation tool.

Many blogs offer syndicated content in the form of RSS (Real Simple Syndication) feeds. For the user who reads many blogs, RSS offers a more manageable experience. RSS feeds can be subscribed to and then aggregated in a single place using any of a variety of RSS aggregation software packages such as Google Reader. These packages not only offer a single point of browsing, such that the user doesn't have to manually visit each blog that they would like to view, but also replaces the disorientation of navigating through a haphazardly connected, or disconnected web of hyperlinks with a structured, easily navigable interface. The benefits of RSS aggregation only apply to content which the user has already discovered and subscribed to. They offer no benefits for navigating a collection of blogs which the user is not familiar with.

4.2. Navigation Tool

The data for the navigation tool was gathered from the world wide web by crawling the search results for a selection of terms related to travel in Thailand. Searches for thai travel, thailand travel, bangkok travel, phuket travel, chiang mai travel, chiangmai travel, and ko samui travel were performed. Each search was limited to a selection of blog hosting sites: blogspot.com, livejournal.com, travelblog.org, travelblogs.com, and travelpod.com. Content cloud data was then generated based on the contents of each URL in the search results.

The navigation tool is a lightweight front-end to this data set. The single HTML page consists of an IFRAME element in which browsed blog sites are loaded into, and the navigation interface. AJAX calls are used to update the navigation interface, and populate the IFRAME element.

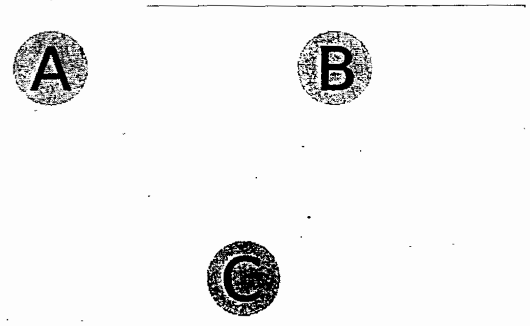


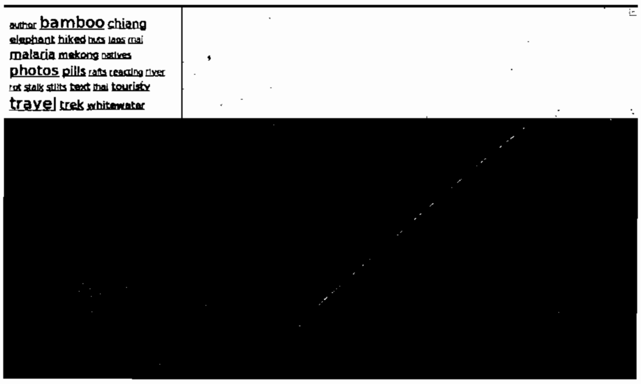
Fig. 3 – Layout of Navigation Tool

The interface provided by the blog navigation tool is divided into three areas. Section (A) displays a content cloud for the currently viewed blog article. Section (B) contains a list of articles related to the currently selected term from the content cloud. Section (C) is where the currently viewed blog is displayed.

Navigation occurs along two axis. When a user clicks on a content cloud term, the list of related articles is updated accordingly. When an article from the list is clicked the content cloud data is updated with terms for the article, and the article is displayed. In this way the user can nagivate though the blog entries in the collection along a path of common terms found in their respective content clouds.

4.3. Tool in Action

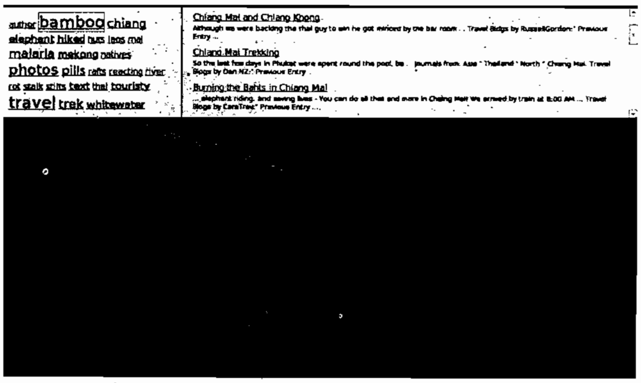
The initial screen displays the content cloud for a document selected from the collection.



author bamboo chiang elephant hiked huts laos mai malaria mekong natives photos pills rafts reaching river rot stalk stilts text thai touristy travel trek whitewater

Fig. 4 – Navigating with Content Cloud

In the following screen shot the user has selected the term 'bamboo' from the content cloud. A selection of documents relevant to bamboo are displayed in the upper right. The selected term in the content cloud is highlighted to inform the user what documents are displayed in the list



author bamboo chiang elephant hiked huts laos mai malaria mekong natives photos pills rafts reaching river rot stalk stilts text thai touristy travel trek whitewater

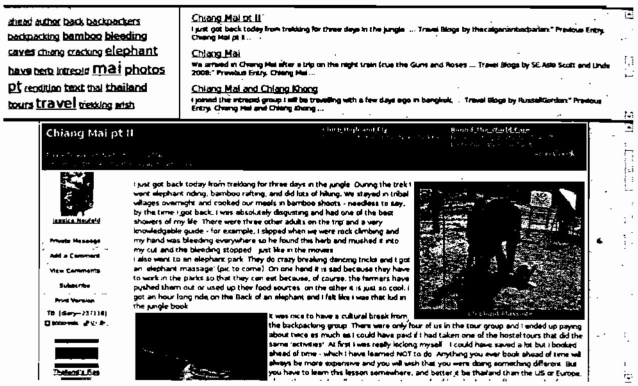
Chiang Mai and Chiang Khong
Although we were backing the thai guy to win he got minced by the bar room . . Travel Blogs by RussellGordon:" Previous Entry ...

Chiang Mai Trekking
So the last few days in Phuket were spent round the pool, be . . journals from: Asia " Thailand " North " Chiang Mai. Travel Blogs by Dan NZ:" Previous Entry .

Burning the Bahts in Chiang Mai
... elephant riding, and saving lives - You can do all that and more in Chaing Mai! We arrived by train at 8:00 AM ... Travel Blogs by CaraTrek" Previous Entry ....

Fig. 5 – Selection of Related Blogs

Once the user clicks on a link in the list, the blog article is loaded into the lower portion of the interface. The content cloud in the upper left hand corner is then updated to contain terms for the currently viewed article. The Article title is highlighted to inform the user which article in the list they are currently viewing.



ahead author back backpackers backpacking bamboo bleeding caves chiang cracking elephant have herb intrepid mai photos pt rendition text thai thailand tours travel trekking wish

Chiang Mai pt II
I just got back today from trekking for three days in the jungle ... Travel Blogs by thecalgarianbarbarian:" Previous Entry. Chiang Mai pt II ...

Chiang Mai
We arrived in Chiang Mai after a trip on the night train (cue the Guns and Roses ... Travel Blogs by SE Asia Scott and Linda 2008:" Previous Entry. Chiang Mai ...

Chiang Mai and Chiang Khong
I joined the intrepid group I will be travelling with a few days ago in bangkok, . Travel Blogs by RussellGordon:" Previous Entry. Chiang Mai and Chiang Khong ...

Chiang Mai pt II

I just got back today from trekking for three days in the jungle. During the trek I went elephant riding, bamboo rafting, and did lots of hiking. We stayed in tribal villages overnight and cooked our meals in bamboo shoots - needless to say, by the time I got back, I was absolutely disgusting and had one of the best showers of my life. There were three other adults on the trip and a very knowledgable guide - for example, I slipped when we were rock climbing and my hand was bleeding everywhere so he found this herb and mushed it into my cut and the bleeding stopped just like in the movies
I also went to an elephant park. They do crazy breaking dancing tricks and I got an elephant massage (pic to come). On one hand it is sad because they have to work in the parks so that they can eat because, of course, the farmers have pushed them out or used up their food sources. on the other it is just so cool. I got an hour long ride on the Back of an elephant and I felt like I was that kid in the jungle book
It was nice to have a cultural break from the backpacking group. There were only four of us in the tour group and I ended up paying about twice as much as I could have paid if I had taken one of the hostel tours that did the same 'activities'. At first I was really kicking myself. I could have saved a lot but I booked ahead of time - which I have learned NOT to do. Anything you ever book ahead of time will always be more expensive and you will wish that you were doing something different. But you have to learn this lesson somewhere, and better it be thailand than the US or Europe.

Fig. 6 – Whole Blog Navigation Tool

Further browsing through the document collection consists of moving between two actvities: selecting documents to view, and seleting terms from the content cloud. Each subsequent selection from the displayed content cloud presents the user with a focused subset of the document collection to browse. The content cloud itself serves as means of navigting between these overlapping sets.

5. EVALUATION

The software displayed here has two distinct components. One piece is the content cloud generation software. The second is a navigation tool which is an application of the content clouds.

5.1 Content Clouds

Content clouds go some way to alleviate the requirements of a tagged folksonomy: that there be a willing community to tag content, and that the community be skilled enough to generate quality tag sets. At the same time we found that content clouds complement the descriptive characteristics of tag clouds. The two key benefits that Content Clouds offer, over Tag Clouds, are firstly the removal of potential bias by process automation. Secondly Content Clouds with a rich semantic can be created for individual documents, whereas conventionally individual documents are only attributed a few tags.

If you consider the navigation tool an application of content clouds, we could also consider applying traditional tag clouds. This view highlights a benefit of content clouds. The tags associated with a document are much less rich than a content cloud. Some documents have as little as one term, and rarely more then five or six. Whereas content clouds generally contain quite a number more. The higher level effect is that the subsets of documents available through a Tag Cloud navigation tool would have less connectivity. In some cases some documents or groups of documents may not have any connectivity. Using content clouds the user is offered more varied and perhaps more specific pathways through the document collection.

5.2 Navigation Tool

While blogs have become an important tool for communicating with customers, the means of navigating the blogoshpere, and categorizing its contents are limited. Search allows a user to arrive at a large collection of documents about a given subject, and through advanced search interfaces the user has the ability to limit the result to only blogs. This generally leaves the user with an overwhelming amount of possibilities and a disorienting web of links to follow. There are a selection of blog specific tools that can help out. These were found to have some weaknesses, which the tool described here seeks to improve upon.

The navigation tool offers a structured means of navigation through a set of related documents. In our example we used a set of search results that a user might come across on their own while browsing the web. We will evaluate the tool by way of comparing it against some of the existing navigation aids available. Blog rolls, random blog links, and RSS aggregation function as a means to link blogs together, though not blog entries themselves. By using content clouds to evaluate specific blog articles, a more granular set of relations can be presented. On a larger scale, by making use of search results the blog article relationships encoded in the content cloud are able to span across multiple blog hosting sites, whereas the aids offered by on blog hosting site are only available within that site.

Because content clouds evaluate article content directly, some of the weaknesses blogrolls are reduced. Generally speaking the there will be more thorough coverage of the relationships between documents, and the relations will be free from human bias. It is possible that a person is better equipped to draw those relations, though the output will invariably be less than what an automated algorithm can produce.

6. CONCLUSION

This paper has addressed some issues which have arisen due to increasing quantities of user generated content, brought about through the Web 2.0 era. The primary concern addressed was that of classification, the difficulty of developing an appropriate taxonomy, and the limitations of the folksonomy approach. The paper introduced a new method of creating the popular cloud visualization, in our case using content rather than tags to create a Content Cloud. By using NLP techniques we reduce the potential for bias that comes from tagging, resulting in a different level of abstraction in the model.

The second concern we addressed was navigation. Navigating a continually growing and evolving space is a complex challenge, so we applied the visualization of Content Clouds to a restricted Blogosphere space to demonstrate a novel and effective way of traversing a content space through articles with related content. This navigation tool is a new development which addresses many issues involved in navigation. Further evaluation of the usability of the tool is required in order to fully analyse its effectiveness.

The paper also introduced the challenges of managing quality and controlling the nature of web based content. These are difficult challenges, so future work may also include investigating how they can be addressed.

7. REFERENCES

[1] I. Assenov and K. J. Cosh, "Destination Branding Evaluation Through Natural Language Processing", *proceedings of Asia Pacific Tourism Association Conference,* 2008.

[2] M.J. Carnot, B. Dunn, A.J. Canas, P. Gram, J. Muldoon, "Concept maps vs. web pages for information searching and browsing", *Institute for Human and Machine Cognition, vol. 2002,* 2001.

[3] K.J. Cosh and P. Sawyer, "Aiding Semiotic Analysis using Natural Language Processing Tools" *in the Proceedings of the 6th International Workshop on Organisational Semiotics* pp 257-269, 2003.

[4] R. Godwin-Jones, "Tag Clouds in the Blogosphere: Electronic Literacy and Social Networking", *Language Learning and Technology,* Vol. 10 No. 2, pp 8-15, 2006.

[5] T. Gruber, "Ontology of Folksonomy: A Mashup of Apples and Oranges", *International Journal on Semantic Web and Information Systems.* Vol. 3 No. 1, pp 1-11, 2007.

[6] M.J. Halvey and M.T. Keane, "An Assessment of Tag Presentation Techniques", *Proceedings of the 16th International Conference on World Wide Web,* pp 1313-1314, 2007.

[7] D. Y. W. Lee, "Genres, registers, text types, domains and styles: clarifying the concepts and navigating a path through the BNC jungle." *Language Learning & Technology*, Vol.5(3): 37-72, 2001.

[8] C. Lindahl and E. Blount, “Weblogs: Simplifying Web Publishing”, *IEEE Computer, pp114-116,* November 2003.

[9] A. Mathes, “Folksonomies: Cooperative Classification and Communication Through Metadata”, *Computer Mediated Communication,* LIS590CMC, 2004.

[10] B. A. Nardi, D. J. Schiano, M. Gumbrecht and L. Swartz, “Why We Blog”, *Communications of the ACM,* Vol. 47, No. 12, pp. 41-46, 2004.

[11] M. Otter, and H. Johnson, “Lost in hyperspace: metrics and mental models” *Interacting with Computers, 13, 1-40. (1),* 2000.

[12] T. O’Reilly, “What is Web 2.0, Design Patterns and Business Models for the Next Generation of Software”, available at www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html (accessed 21 May 2008) 2005.

[13] C. S. Peirce, *Collected Writings* (8 Vols.). Hartshorne, C., Weiss, P. and Burks, A.W., (Eds.). Harvard University Press, 1931-1958.

[14] P. Rayson, L. Emmet, R. Garside and P. Sawyer, “The REVERE Project: Experiments with the application of probabilistic NLP to Systems Engineering.” *In Proceedings of NLDB ’00, the 5th International Conference on Applications of Natural Language to Information Systems,* 2000.

[15] A.W. Rivadeneira, D. Gruen, M.J. Muller and D.R. Millen, “Getting Our Heads in the Clouds: Towards Evaluation Studies of Tag Clouds”, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* pp 995-998, 2007.

[16] Z. Shevked and L. Dakovski, “Blogging – A Modern Paradigm in Internet Communication Technologies”, *In JVA International Symposium on Modern Computing,* 2006.

[17] F. Saussure, *Course in General Linguistics* (trans. Wade Baskin). Fontana/Collins, London, 1974.

[18] N. L. Waters, “Why You Can’t Cite Wikipedia in My Class”, *Communications of the ACM, Vol 50, No. 9,* September 2007