Deep context with a sense-of-self

Robert McArthur (CSIRO ICT Centre)

TREC Blog Track

Introduction

There is a trend in information retrieval to an increased involvement of *context*. The success of recent workshops at SIGIR [8] and the subsequent 2006 IIiX Symposium[1], the SIGIR exploratory search workshop[2], and the outcome of the 2004 SWIRL workshop[3] [21] are indicators that elements of user context are both sort and becoming available. *"The provision of tools…can yield great rewards for users, especially when contextual factors such as user emotion, task constraints, and dynamism of information needs are considered"* [26].

Recent weblog workshops[4] reflect a strong interest in the discovery of context about the user in the form of the blogger's age [2, 24], gender [23, 24], opinions, sentiments and opinions expressed [9, 19, 22], mood levels [1, 20], happiness [18], residential area [28] and social network(s) [6, 7, 11-13, 25]. The research concentrated on particular forms of identifying or extracting the information, providing fertile ground for TREC-style comparisons of the approaches.

Considerable interest has been shown in the analysis of sentiment in weblogs ([1, 4, 18-20, 22]). This broad umbrella encompasses notions of mood, opinion, emotion and happiness. Sentiment is only one aspect of user context. Indeed, many of the published notions of sentiment derive from reflection of authorship. That is, determination of the mood, opinion or emotion comes through analysis of the artefacts of communication, the blog entry, and is deemed a reflection of the sentiment of the author *at the time*. Of course, the assumption is that a person's writing reflects their inner state of being. Although it is tempting to, repeating history, ignore the author and concentrate on the blog entry itself, the focus should remain on determining more about the user.

To this end, more information about the user is required apart from the emphasis on the (usually) explicit manifestations of their selves. Figure 1 presents this difference by separating what we know about the document (e.g. style), what we know about the user from the document at a particular point in time (e.g. sentiment), and the more implicit and meta-information which is derived from the sentiment and which captures deeper context about the user. An example of such information is a computational manifestation of a person's sense-of-self. This has been investigated previously using mailing list records in an online health setting ([15]) and this paper uses it as an exemplar of deeper user context, showing how to apply the ideas and techniques to blog data in a TREC setting.

---

[1] http://www.db.dk/IIiX/

[2] http://research.microsoft.com/~ryenw/eess/index.html

[3] http://www.cs.mu.oz.au/~alistair/swirl2004/

[4] 2006 AAAI Spring Symposia on Computational Approaches to Analysing Weblogs, http://www.umbriacom.com/aaai2006_weblog_symposium/ and 3rd Annual Workshop on the Weblogging Ecosystem (WWW2006), http://www.blogpulse.com/www2006-workshop/
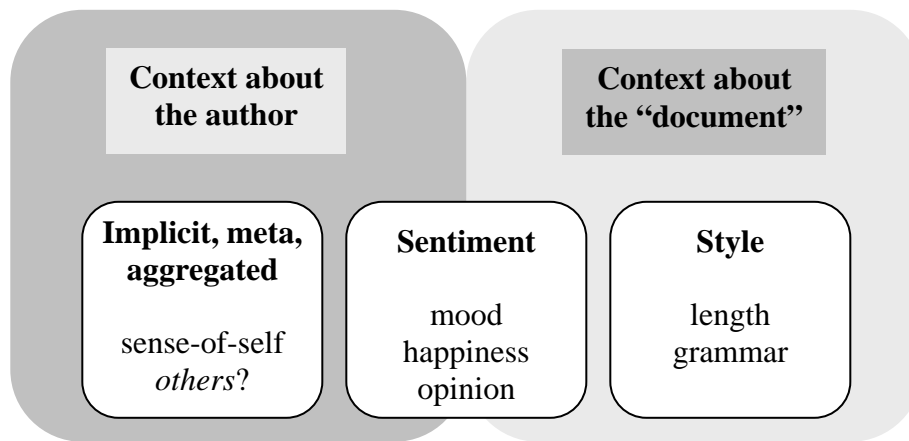
Figure 1: Different contexts, from document to user

Description

This paper describes one approach to identifying implicit or aggregated user context. Such context can be useful in information retrieval in various ways. For example, credible identification of people with a mental illness who are tending towards suicide can be a tool for clinical intervention. More directly in the sphere of IR, a general sense-of-self could be combined with an expertise identification system to rank blog entries in a standard search so that entries by someone both confident *and* knowledgeable are ranked more highly.

The particular method used to identify a person's sense-of-self follows directly from previous work ([15]): a socio-cognitively motivated ([5, 27]) technique, hyperspace analogue to language (HAL; [3]), is used to create a semantic space. Certain terms, chosen to exemplify the concepts of negative emotion and kin, are combined to form a 2D space onto which the monthly sense-of-self vectors of the authors are projected. The sense-of-self vectors arise from the identification of first-person language within the blog posts. A large movement within the 2D space of the monthly sense-of-self points indicate a potentially interesting change to the author's sense-of-self. Terms used to create the axes are drawn from the original work, as well as more recent blog-related sources (like [18]).

Application to blogs

Identification of user context is difficult. The common formal information retrieval paradigm ignores the author of documents, apart, perhaps, from their name. However, blog entries, unlike many types of documents, are written by the individual as a personal communication. Sometimes they are written in the knowledge that many thousands of people will read, digest and comment on the contents. Other blogs are more personal and it doesn't seem to matter to the author whether *anyone* else reads them. Whichever is the case on a connection through these poles, the blogs have a clear authorship and are often written in first person. Apart from email, which is notoriously difficult to conduct research on because of privacy concerns, blogs are the main other medium which has these features, and it is these features which provide the strongest source of information – context – about the author.

In summary, the advantage of blogs as against all other forms of data for identifying and analysing user context is that

1. they are written by a single, identifiable author;
2. they are revealing – often explicit personal information is presented;

3. they are freely available.

Methods

The weblogs provided in 2006 were analysed in a manner similar to [15]. Unlike in that study, some "pre-semantic" information such as part-of-speech tagging was not derived as it was not required because of the amount of data ([14] describes further motivations); for similar reasons, SVD (singular value decomposition) was not applied, hence only explicit associations or relationships were uncovered.

To further aid analysis for these particular experiments, 9 consecutive days in the first part of the data, December (20051207, 20051208, 20051209, 20051210, 20051211, 20051212, 20051213, 20051214, 20051215), and 7 days in the latter period of February (20060201, 20060202, 20060203, 20060204, 20060205, 20060206, 20060207) were analysed. Multiple days in each period were chosen to attempt to capture one of the qualities of personal blogging – that repeated daily or near-daily entries were the type that are more likely to evince and respond to the deep personal analysis of the type this research describes. The differences between the data from both ends of the period available was chosen since evidence of a deeper personal change is likely to be more pronounced over a longer period; as well, it restricted the amount of data to a more manageable level for the semantic space analysis.

In summary, the following stages were implemented, with comments in italics underneath:

1. The text of the Blog entries were extracted from the XML-like syntax;
   This was more difficult than necessary due to the XML-like format used, the fact that the messages were often in their original HTML format – each of which used different methods and standards including Javascript – and the convolutions with differing character sets involved.

2. Entries were tokenised, and non-English language blogs eliminated;
   This again was difficult as the texts were very 'dirty' – many strange characters, odd positioning of spaces and inter-word separators. It is unlikely more detailed linguistic analysis, such as part-of-speech and chunking, would succeed well because of both the format of the data as well as the language used. Non-English blogs were relatively easy to identify and remove using simple regexps.

3. A small number of stopwords were removed;
   Some words usually considered stopwords, like "I" or "my", are words that are vital to understanding the deep personal information in the blog. Removal of too many words that are important indicators in terms of personal context risks loss of the ability, wholly or in part, to find the desired entries; the cluttering of the vectors with 'useless' words harms the quality of the subsequent analysis. The choice of which and how many stopwords to remove is likely to differ depending upon the particular personal context under examination.

4. General cleaning of the text was performed (translation of forms like "I've" to "I have", removal of URI's, converting to lower-case)

5. Hyperspace analogue to language ([3]) was performed on the tokenised text over all the blog entries identified above. The window size was chosen at 10 (as per [3]). The result was the creation of two combined vectors, one for the set of *kin* words and one for the set of *negative emotion* words.
   Normally a semantic space is constructed for all words in the corpus. However, due to the large amount of data, only vectors for terms of interest were computed rather than

an entire semantic space (a vocabulary of 1,054,820 terms was identified; 556074 were not known to Unix spell leaving 498746 known terms; even this is a very large semantic space compared to [3] or [10]). Terms of interest are shown in Table 1. *In keeping with previous research [15-17], pre- and post- associations were combined leaving a single vector for each word. All vectors were normalised to unit length. Each of the two vectors representing sets of words were created by summing the individual word's vectors, averaging and then normalising the result.*

6. Hyperspace analogue to language ([3]) was performed on the tokenised text (window size=10) but only vectors for the sense-of-self were created. Also, since the sense-of-self is personal, the sense-of-self vector (calculated as in the above bullet point) was created separately for each blog feed over the days in question. *Over 56,000 sense-of-self vectors were created, with over 13,300 appearing on more than one day during the analysis period. Those occurring on only one day were eliminated.*

7. The sense-of-self vectors were projected onto the two axes of interest using simple vector algebra: each sense-of-self vector for a blog feed in a day (vector of 1 x *n*) was projected onto the combined *kin* vector (1 x *n*) and the combined *negative emotion* vector (1 x *n*) yielding an *x,y* position against these two axes. *The length of the sense-of-self vectors varied widely – the more contexts in which the terms comprising the sense-of-self (*Table 1*) were used, the larger the vector. A larger vector* may *be a better exemplar of the person's sense-of-self, although it is likely to be a non-linear relationship – one occurrence of "I" may not be strongly evidential, but three occurrences may be as good as 30.*

Negative emotion words: "hate", "pain", "anger", "painful", "fatigue", "fatigued", "angry"
Kin words: "mother", "father", "fathers", "grandfather", "mothers", "grandmother", "mother-in-law", "grandson", "sons", "son-in-law", "daughters", "granddaughter", "daughter-in-law", "sibling", "siblings", "husbands", "sister", "brother", "daughter", "son", "wife", "husband"
Words comprising the sense-of-self: "I", "me", "my", "mine"

Table 1: Terms of interest

While not required output for TREC-style analysis, Table 2 presents the largest associations in the combined *kin* and *negative emotion* vectors to assist understanding of the method. For ease of reading, the normalised values (0-1) have been multiplied by 100,000. Bold indicates terms comprising the combined vector; italics indicates "interesting" associations.

Kin		Negative emotion	
(n=123,589, x̄ =10.23)		(n=69,442, x̄ =20.04)	
My	57423	i	68227
i	38471	not	21778
his	24290	my	19564
a	18530	me	17689
her	18476	you	15234
me	16860	but	14084
possessive	15774	have	13189
you	14362	back	12397
not	13014	so	11483
he	12527	a	11427
personalpossessive	12478	*ellipsis*	*11382*
have	11745	**hate**	**10481**
your	10595	**pain**	**10370**
she	9359	they	10322
had	8187	do	10219
mother	**7009**	them	8125
but	6450	*all*	*7986*
one	5657	*like*	*7756*
we	5330	am	7739
so	5271	is	7586
has	5251	he	7490
sex	*5028*	*love*	*7222*
him	5003	just	6709
ellipsis	*4845*	much	6340
their	4729	it	6128
wife	**4722**	your	5992
like	*4626*	up	5918
am	4604	*more*	*5781*
up	4568	his	5711
sister	**4492**	we	5589
they	4408	one	5434
our	4366	*people*	*5262*
do	4337	her	5126
said	*4294*	she	5115
father	**4170**	get	5065
was	4162	really	5034
all	4070	*because*	*4899*
now	4066	did	4857
just	3805	there	4733
would	3788	him	4611
time	*3730*	*feel*	*4525*
little	*3690*	their	4250
son	**3618**	out	4230
is	3582	*management*	*4105*
brother	**3489**	possessive	4065
nude	*3478*	know	3958
out	3403	being	3892
incest	*3386*	**anger**	**3803**

Table 2: Largest associations in the combined kin and negative emotion vectors

Example

The following table (Table 3) shows an example ranked list of feed numbers. The score indicates the level of difference (Euclidean distance) between the 2D sense-of-self in one day compared to another day; it is the Euclidean distance between the 2D point (*kin* axis, *negative emotion* axis) on one day compared with another day. The vector size is also shown, along with the number of elements of the sense-of-self vector compared with the combined *kin* and *negative emotion* vectors.

Score	Blog Feed ID	*Kin* value	*Negative emotion* value	Day	Vector size	No. elements in common	
						with *kin*	w/*negative emotion*
0.8660	BLOG06-feed-021858	0.0533	0.0521	20051215	16	16	16
		0.5871	0.7136	20060202	776	737	716
0.7868	BLOG06-feed-074342	0.1892	0.2010	20060202	29	28	28
		0.6470	0.8409	20051215	776	750	740
0.7708	BLOG06-feed-036441	0.1284	0.1811	20051209	31	31	31
		0.6391	0.7584	20060203	911	871	852
0.7577	BLOG06-feed-020809	0.1552	0.2000	20051208	42	37	37
		0.6589	0.7743	20060202	1286	1191	1180
0.7528	BLOG06-feed-073897	0.1705	0.1933	20051215	37	36	36
		0.6431	0.7793	20060202	749	729	714
0.7397	BLOG06-feed-036157	0.0978	0.0712	20060203	29	28	27
		0.5780	0.6338	20051209	473	448	437
0.7280	BLOG06-feed-024861	0.1467	0.1527	20051215	43	42	42
		0.6044	0.7188	20051208	572	546	538
0.7127	BLOG06-feed-032182	0.0156	0.0351	20051209	26	12	12
		0.4542	0.5968	20060203	57	57	56
0.6967	BLOG06-feed-073960	0.1477	0.0803	20051215	47	45	41
		0.5583	0.6432	20060202	609	571	551
0.6821	*BLOG06-feed-032140*	*0.2250*	*0.2927*	*20051209*	*221*	*213*	*205*
		0.6898	*0.7919*	*20060203*	*725*	*678*	*681*
0.6814	BLOG06-feed-035775	0.0044	0.0098	20051209	11	11	11
		0.4542	0.5216	20060203	78	77	76
…	…	…	…	…	…	...	…

Table 3: Example output (*n*=9230 pairs, avg. score=0.10)

Assessment

The assessment as to which blog entries most contribute to the context is almost certainly qualitative and task dependent, in line with the socio-cognitive nature of the problem. In the example of the mental health care worker, the determination of whether a particular blog entry *is* the basis for intervention is there own; possibly their peers would disagree with their conclusion. That's the bad news; the good news is that often the identification is quite viable since as humans we do such comparisons every moment and are, relatively, good at it.

For example, in the example shown in Table 3, let us examine the particular blog entries that form the basis of one of the results. The list is ordered by a basic similarity

score, however, a factor not encompassed by that score is the size of the sense-of-self vector. A small vector means few uses of first person language and hence may not be reliable. Since in the example the second-last entry (BLOG06-feed-032140) has the largest number of associations in the two day's sense-of-self vectors, it will serve as a practical case in point.

The first day for this person (aka individual blog feed) is 20051209, where a sense-of-self is (*0.2250, 0.2927*), while the second is 20060203 (*0.6898, 0.7919*). Instances of blog entries from the first day's feed are presented below. In bold is the title followed by the unadorned entry. Blog entries that are similar to, or do not add to those shown, are not presented (there were 10 entries in total).

The 12 Days of Thanks: Day 10   Things I am thankful for #10: Bruce Campbell   This... is my BOOMSTICK!!!   One of the best cheesy actors of our time, the star of the Evil Dead series. I simply can't resist a movie with Bruce.
The 12 Days of Thanks: Day 8   Things I am thankful for #8: Swearing   ...or as I like to call it, "Fluent Second Language"
Intermission   I am taking a brief break from giving thanks to express a few thoughts. I think I will commission a movie about myself, and allow the role of ME to go to Salma Hayek. It's eerie--it's like she's my twin. Gives me shivers... The only problem is that she'd have to train for months to learn how to be as dedicated to pure laziness as I am. It would mean months and months of sitting on the same spot on the couch. I'm not sure Salma is that good of an actress.   Koala, the fifth kitten in the litter I've mentioned a half a million times disappeared for a while. I couldn't figure out why. Until last night. We have a fucking POSSUM living in our crawlspace. No wait, let me correct that. We have a fucking POSSUM living in our BACK crawl space. You see, we have two crawlspaces. One in the front and one in the back. The one in the front is occassionally squatted by a homeless CRACK junkie. It's like having a tenant who shits the floor and smells of whiskey-soaked rotten eggs, but won't pay the rent. Sort of like a teenager. Eventually we will buy a new set of hardware for the hatch door to keep him out... or keep him IN, depending on whether or not I've taken my Maca that day. No idea what we're going to do about the GIANT rodent in the back crawlspace. So basically, Koala the kitten has sought residence with his Daddy, whom I've only seen once, but looks like a bigger version of Booga. Booga's going to be a big boy. He's going to whip Mojo's ass one day. And now I return you the regularly scheduled Days of Thanks...

Example blog entries from the second day's (20060203) feed are (10 entries total):

Cutting Day   Date: February 24, 2006   5:30AM   Event: Laparoscopy/Hysteroscopy/Septum Eviction 2006   Prep: Occasional pacing; random anxiety dream; shallow breathing; temporary attention deficit; excitement and hope that this might fix the plumbing problems. Three weeks to go.....
Ahem... yeah.   Apparently, my MRI strongly suggests an arcuate uterus. Arcuate, as in irregularly shaped. Arcuate, as in almost heart-shaped (and not in a cheeky Valentines kind of way). Arcuate, as in what one doctor noted when I was pregnant, only to be contradicted by Dr. Asshole later. Arcuate, as in lots more surgery. Arcuate, as in "might even be septate--but we won't be sure until the surgery because there's a gigantic fibroid blocking our view." I am having a difficult time finding enough information about it. Most say that it doesn't increase the risk of miscarriage, which I have to say is royal bullshit according to my interview of my seven deceased fetuses. The Dr. Google search is afoot. A funny tidbit.. when you type "arcuate uterus" into WebMD.com, it replies, "Don't you mean ADEQUATE UTERUS?" WebMD, you are a smart ass.

They like it when you sweat...   Nothing yet. All day, I waited patiently by the phone. I took my cell phone with me to staff meetings, expecting to leave the meeting to talk to Dr. Awesome. ...and I waited.... ...and waited... ...and waited... At 5:55pm, I got up and headed out the door to go home. Somewhere en route from my desk to the car, they called. But of course, the windtunnel I walk through known as "parking deck" was noisy enough that OF COURSE, I didn't hear the damn phone ringing. Ah, but they left me a message. "Mrs. Drab, we have the results back for your MRI. If you will call us tomorrow, we will go over those results with you. Also, you need bloodwork drawn. Call and we will give the details." Thanks, that was a fat load of help.   My imagination is running wild. Now, I'm fairly certain THIS is The Shadow. This is why they are being coy. They don't want to tell me about the evil creature that has attached itself to my uterine wall. MMMMhmmmm... Mark my words. E-V-I-L C-R-E-A-T-U-R-E. Closely related to the Ripapod . Dammit, they found a mutant Ripapod in there. Son of a bitch.

Why I will fail at Anger Management   My doctor (the one who treated my TMJ) advised me to work on my anger management issues. I will fail at this task. It's not that I'm a defeatist or like to give up easily (anyone who's followed my ridiculous Chronicles of Conceiving knows that). The truth is that I've had this anger as long as I can remember. It's like a close friend. The one who leaves cigarette burns on your couch and never replaces the toilet paper, but at the end of the day you are saying "OHHHH Anger. You're simply ADORABLE." I'd miss the little curmudgeonette if she were to leave me entirely. And let's face it, with an asshole around every corner, it's a virtual impossibility. For example, Gwyneth. She took a lovely little band like Coldplay and twisted it with her evil witchy magic. Here's the transition: BEFORE (c.2000)   I awake to see that no one is free We're all fugitives Look at the way we live Down here, I cannot sleep from fear, no I said, which way do I turn Oh, I forget everything I learn...   AFTER (c.2005) You cut me down a tree and brought it back to me And thatÃ<83>ÃÂ¢Ã<82>Ã<8>Ã<82>Ã<99>s what made me see where I was going wrong You put me on a shelf and kept me for yourself I can only blame myself, you can only blame me...   Do you SEE what she did? She turned Coldplay into a teenage girl writing shitty poetry about ponies and feelings! How can I not be angry about that? But then I see this, and it helps me to overcome my anger. Look at that Glad trash bag she has around her neck! How can I be mad when Gwyn's prancing about in a garbage bag? It's hilarity at its finest. Chris Martin has been bewitched by a bag lady! But there are plenty of other assholes where Gwynnie drops off. It will always be a challenge, and I'm not so sure I'm ready to give it up. Thank GOD for Vicodin, huh?

Clearly, there is a difference between how and what is being expressed in the two sets of blog entries, and in particular what is "behind" that expression – the first day's entries portray someone who is outward looking, while the second is much more inwardly focussed and personal. Although the use of first-person in the two sets of entries has not substantially changed, the nature of the messages is indicative of a change in sense-of-self (using the same understanding of the particular meaning of this concept from [15]).

To return to a more general TREC sense, he above example is a strong indication that a change in deeper user context is feasibly identifiable. Assessors would need to annotate the change to indicate their belief in the strength of difference in the context over the period of time. The guidelines given to assessors would also need to be suitably clear, while allowing the human judgement that is clearly required to be exercised in order to assess the context. Importantly, and in contrast to the usual TREC beliefs, inter-assessor differences should perhaps be embraced rather than ameliorated, with the annotated explanations forming a vital part of the assessment and evaluation.

TREC Evaluation metrics

The approach to evaluation for TREC for the implicit user context, is to identify the particular blog entries which contribute to the context. In the case of this exemplar of a computational sense-of-self, the first activity is to identify those blog authors whose sense-of-self is substantially changing. The evaluation then proceeds to list and rank the particular blog posts that most contributed to the change in sense-of-self. To place this evaluation in a task perspective, a mental health care worker is performing qualitative analysis of blog postings; rather than having to peruse all postings, this evaluation allows the retrieval of a small subset of postings for detailed qualitative analysis, e.g., the identification of an intervention point.

Conclusion

The current TREC Blog track opinion task is an interesting example of capturing user context. The contention of this paper is that while other sentiments can likewise be mined, such as mood and happiness, a deeper understanding of an author can lead to useful context about the person. This context cannot be derived or discovered from the impersonal documents often used for evaluation; implicit, aggregate information about the author must come from the author's own personal descriptions. The next generation of search systems will make more use of context, desiring to go deeper into the person and personal and what "makes us tick". A computational sense-of-self is but one element of this context.

Bibliography

1.   Balog, K. and Rijke, M.d., Decomposing Bloggers' Moods: Towards a Time Series Analysis of Moods in the Blogosphere. in *Third annual workshop on the {Weblogging} ecosystem, WWW2006*, (Edinburgh, 2006).
2.   Burger, J.D. and Henderson, J.C., An Exploration of Observable Features Related to Blogger Age. in *AAAI Spring Symposia 2006 on Computational Approaches to Analysing Weblogs*, (Stanford, California, 2006).
3.   Burgess, C., Livesay, K. and Lund, K. Explorations in context space: words, sentences, discourse. *Discourse Processes*, *25* (2\&3). 211-257.
4.   Cesarano, C., Dorr, B., Picariello, A., Reforgiato, D., Sagoff, A. and Subrahmanian, V.S., OASYS: An Opinion Analysis System. in *AAAI Spring Symposia 2006 on Computational Approaches to Analysing Weblogs*, (Stanford, California, 2006).
5.   Gärdenfors, P. *Conceptual Spaces: The Geometry of Thought*. MIT Press, 2000.
6.   Gu, L., Lento, T., Smith, M. and Johns, P., How do blog gardens grow? Language community correlates with network diffusion and adoption of blogging systems. in *AAAI Spring Symposia 2006 on Computational Approaches to Analysing Weblogs*, (Stanford, California, 2006).
7.   Hsu, W.H., Weninger, T., Pydimarri, T. and Paradesi, M.S.R., Collaborative and Structural Recommendation of Friends using Weblog-based Social Network Analysis. in *AAAI Spring Symposia 2006 on Computational Approaches to Analysing Weblogs*, (Stanford, California, 2006).
8.   Ingwersen, P. and Järvelin, K. Information retrieval in context: IRiX. *SIGIR Forum*, *39* (2). 31-39.
9.   Ku, L.-W., Liang, Y.-T. and Chen, H.-H., Opinion Extraction, Summarization and Tracking in News and Blog Corpora. in *AAAI Spring Symposia 2006 on Computational Approaches to Analysing Weblogs*, (Stanford, California, 2006).
10.   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*. 211-240.

11.   Lento, T., Welser, H.T., Gu, L. and Smith, M., The Ties that Blog: Examining the Relationship Between Social Ties and Continued Participation in the Wallop Weblogging System. in *Third annual workshop on the {Weblogging} ecosystem, WWW2006*, (Edinburgh, 2006).

12.   Li, X., Liu, B. and Yu, P.S., Mining Community Structure of Named Entities from Web Pages and Blogs. in *AAAI Spring Symposia 2006 on Computational Approaches to Analysing Weblogs*, (Stanford, California, 2006).

13.   Lin, Y.-R., Sundaram, H., Chi, Y., Tatemura, J. and Tseng, B., Discovery of Blog Communities Based on Mutual Awareness. in *Third annual workshop on the {Weblogging} ecosystem, WWW2006*, (Edinburgh, 2006).

14.   McArthur, R. Computing with meaning by operationalising socio-cognitive semantics, PhD, *Faculty of Information Technology*, Queensland University of Technology, Brisbane, 2006.

15.   McArthur, R., Bruza, P., Kralik, D. and Warren, J., Projecting computational sense of self: A study of transition in a chronic illness online community. in *Proceedings of the 39th Hawii International Conference on System Sciences (HICSS-39)*, (2006).

16.   McArthur, R. and Bruza, P.D., Discovery of implicit and explicit connections between people using email utterance. in *Proceedings of the 8th European Conference on Computer-supported Cooperative Work (ECSCW)*, (2003), Kluwer Academic Publishers, 21-40.

17.   McArthur, R., Bruza, P.D. and Song, D., Policy Conformance in the Corporate Blog Space. in *Workshop on Policy Management on the Web, WWW'05*, (2005).

18.   Mihalcea, R. and Liu, H., A Corpus-based approach to finding happiness. in *AAAI Spring Symposium on Computational Approaches to Weblogs*, (2006).

19.   Mishne, G. and Glance, N., Predicting Movie Sales from Blogger Sentiment. in *AAAI Spring Symposia 2006 on Computational Approaches to Analysing Weblogs*, (Stanford, California, 2006).

20.   Mishne, G. and Rijke, M.d., Capturing Global Mood Levels using Blog Posts. in *AAAI Spring Symposia 2006 on Computational Approaches to Analysing Weblogs*, (Stanford, California, 2006).

21.   Moffat, A., Zobel, J. and Hawking, D. Recommended Reading for IR Research Students. *SIGIR Forum*, *39* (2).

22.   Mullen, T. and Malouf, R., A preliminary investigation into sentiment analysis of informal political discourse. in *AAAI Spring Symposia 2006 on Computational Approaches to Analysing Weblogs*, (Stanford, California, 2006).

23.   Nowson, S. and Oberlander, J., The Identity of Bloggers: Openness and gender in personal weblogs. in *AAAI Spring Symposia 2006 on Computational Approaches to Analysing Weblogs*, (Stanford, California, 2006).

24.   Schler, J., Koppel, M., Argamon, S. and Pennebaker, J., Effects of Age and Gender on Blogging. in *AAAI Spring Symposia 2006 on Computational Approaches to Analysing Weblogs*, (Stanford, California, 2006).

25.   Thelwall, M., Blogs During the London Attacks: Top Information Sources and Topics. in *Third annual workshop on the {Weblogging} ecosystem, WWW2006*, (Edinburgh, 2006).

26.   White, R.W., Kules, B., Drucker, S.M. and schraefel, m.c. Supporting exploratory search: introduction to special issue. *Commun. ACM*, *49* (4). 36-39.

27.   Widdows, D. *Geometry and Meaning*. CSLI Publications, 2004.

28.   Yasuda, N., Hirao, T., Suzuki, J. and Isozaki, H., Identifying Bloggers' Residential Area. in *AAAI Spring Symposia 2006 on Computational Approaches to Analysing Weblogs*, (Stanford, California, 2006).