Policy Conformance in the Corporate Blog Space

R.McArthur
Distributed Systems Technology Centre
Brisbane,Australia
mcarthur@dstc.edu.au

P.D.Bruza
Distributed Systems Technology Centre
Brisbane,Australia
bruza@dstc.edu.au

D.Song
Distributed Systems Technology Centre
Brisbane,Australia
dsong@dstc.edu.au

ABSTRACT

This paper describes part of a solution to the interpretation of human-readable policy documents into semi-automatic conformance checking. Using a socio-cognitively motivated representation of shared knowledge, and applying appropriate inference mechanisms from a normative perspective, a mechanism to automatically detect potentially non-conforming blog entries is detailed. Candidate non-conforming blog entries are flagged for a human to make a judgement on whether they should be published. Analysis of data from a public corporate blog is analysed and results suggest the methodology has merit.

Categories and Subject Descriptors

I.7. [Document and Text Processing], H.4. [Information Systems Applications]

General Terms

Algorithms, Management, Experimentation, Human Factors, Theory

Keywords

Semantic space, policy conformance, blog, knowledge management

1. INTRODUCTION

Managers of organisations have long tried to control what is the official word of the body versus what an employee personally has presented. The (generally) open nature of the WWW has meant an increasing desire for control by some managers, while others have realised the need for a different way of working – for example, the open source software movement. Management in this new way isn't laissez faire, it respects the possibilities of more openness but still has control, often through loosely worded policy rather than the heavy legal jargon. This approach can be characterised as being more carrot than stick.

Sun Microsystems has recently created a standard blog space[1] available to all employees, visible to the world. From Tim Bray's website, on the 6 June 2004[2]:

> *It's been running for some time, and it's stable enough now to talk about in public: blogs.sun.com is a space that anyone at Sun can use to write about whatever they want. The people there now are early adopters; there's an internal email going out to the whole company Monday officially reinforcing that blogging policy, encouraging everyone to write, and pointing them at blogs.sun.com.*

The Sun Policy on Public Discourse[3] is written for people. It encourages blogging stating "*As of now, you are encouraged to tell the world about your work, without asking permission first (but please do read and follow the advice in this note).*" Because of the implications of the policy, and the particularities and importance of wording and intentionality, we have reproduced it in entirety in Appendix A. Appropriate parts are quoted in the following sections.

This paper describes part of a solution to the interpretation of human-readable policy documents into semi-automatic conformance checking. Using a socio-cognitively motivated representation of shared knowledge, and applying appropriate inference mechanisms, a mechanism to automatically detect potentially non-conforming blog entries is detailed. Candidate non-conforming blog entries are flagged for a human to make a judgement on whether they should be published. Figure 1 shows the workflow.

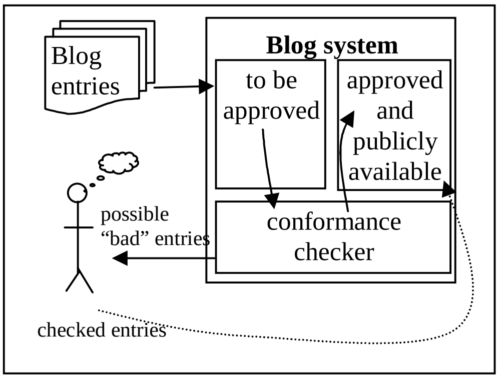

Figure 1: Workflow

The benefits are a significant lessening of work for humans to evaluate each blog entry. Instead, only a subset is required to be vetted by a person.

The remainder of this paper describes the approach taken in more detail, starting with the notion of normative disconformance and applying semantic spaces to blog data,

[1] http://blogs.sun.com/

[2] http://www.tbray.org/ongoing/When/200x/2004/06/06/BSC

[3] http://www.tbray.org/ongoing/When/200x/2004/05/02/Policy (note this was so over the time of this study but may have changed)

thence to experimental results of examining Sun's blog data with respect to one element of its policy.

2. OPERATIONALISING NORMATIVE DISCONFORMANCE

Let N be a normative model comprising principles (or standards) S_1,...Sn. Let B be a piece of augmentative behaviour. Let B disconform with principle Si. If Si is genuinely normative then B is a mistake (at a minimum) [1].

We believe that Sun's problem with checking compliance of blog content can be considered conceptually from a normative perspective. With respect to Sun, read "mistake" as a breach of policy.

Implementing this requires firstly a computational variant of the normative model N, as well as an (semi-) automated procedure for determining (or estimating) disconformance.

Cognitive science distinguishes between three models of cognitive performance:

1. the normative model N that sets standards of rational performance, irrespective of the (computational) cost of compliance;
2. the prescriptive model P which attenuates the standards to make them executable; and
3. the descriptive model D which is a law governed account of actual performance.

Sun's policy can be considered as a high level prescriptive model. It is assumed that the human moderators apply quite some background knowledge B in order to determine or surmise disconformance.

It seems unlikely that a sufficiently large training set of disconforming blog entries can be acquired, therefore a supervised learning approach is almost certainly not appropriate for detecting disconformance. We take a different approach. Certain words, or phrases, in the prescriptive model flag concepts that are key to a particular standard. These can be considered as pseudo-queries with which blog entries can be retrieved and ranked.

It is well known from the field of information retrieval that short queries are typically imprecise descriptions of the associated information need. More effective retrieval can be obtained via automatic query expansion the goal of which is to "guess" related terms to the query at hand. The word "guess" is used deliberately here as the system is ignorant of the actual information need.

Considered in this light, query expansion is a manifestation of abduction. The goal is to abduce related terms to the pseudo-query which are relevant to the intention behind the pseudo-query. If the query expansion mechanism abduces poorly, retrieval precision will decline, a consequence of which is that disconformant blog entries will not be highly ranked in the retrieval ranking. In this article, we will employ a query expansion mechanism which abduces expansion terms by computing the information flow between concepts in a high dimensional semantic space. Query expansion experiments carried out in a traditional information retrieval setting have shown information flow to be promising, particularly for short queries [2].

3. SEMANTIC SPACES

Nonaka and Takeuchi [3] produced an important and viable knowledge creation system in 1995. We have instantiated their notion of an externalisation mode in which tacit knowledge is made explicit and "*The semantic aspect of information* [as against the syntactic] *is more important for knowledge creation, as it focuses on conveyed meaning*.", with Freyd's [4] work on *shareability* which posited

> "*a dimensional structure for representing knowledge is efficient for communicating meaning between individuals. That is, a small dimensional structure with a small number of values on each dimension is argued to be especially shareable, which might explain why such structures are observed*." (Pp.198-9)

The combination of the explicit-tacit knowledge mode with the dimensional representation is further strengthened by Gärdenfors' three level socio-cognitive model of cognition [5]. He argues that meanings of words come from conceptual structures in people's heads – they emerge from the conceptual structures harboured by individual cognition together with the linguistic power structure within the community. Of his three levels of representation, symbolic, conceptual and associationist (sub-conceptual), it is the middle, conceptual, level that is of relevance for this paper.

People write blog entries to communicate. In all communication, there are both explicit and tacit parts to the message. Ducheneaut and Bellotti [6] found that:

> *Persistent talk* [in email] *provides the context for the solitary activity of viewing the content to which it relates…However, during our interviews we, in fact, saw many more examples of imprecise references that were immediately understood than long, drawn-out, explicit and literal descriptions or references.*" and
>
> "*…email conversations are grounded in sufficient mutual understanding to allow very brief, sketchy and implicit references to succeed without posing significant problems in interpretation.*"

Compliance analysis of blog entries with respect to any policy, whether perfectly formed or not, is always dependent on the language used in the entry. Explicit mention of keywords is unlikely to uncover the range of candidate non-compliant entries that make up blog data in the "real world", and will most likely result in poor recall and precision (concepts from information retrieval).

Our previous work [7,8,9] has shown the efficacy of a socio-cognitively based dimensional structure-a semantic space-as a knowledge representation framework. Although there are a number of algorithms for populating such a space, we will briefly describe one, HAL, below. We will then discuss ways of using the semantic space in the context of compliance and blog data.

3.1 Creating the representation - HAL

Hyperspace Analogue to Language (HAL) is a model and technique to populate a semantic space [10,11]. HAL produces vectorial representations of words in a high dimensional space that seem to correlate with the equivalent human representations [12]. For example, word associations computed on the basis of HAL vectors seem to mimic human word association judgments. HAL is "*a model that acquires representations of meaning by capitalizing on large-scale co-occurrence information inherent in the input stream of language*".

Words from communication–blogs–are represented in dimension structures through HAL. The space comprises high dimensional vector representations for each term in the vocabulary. Briefly, given an *n*-word vocabulary, the HAL space is a *n*x*n* matrix constructed by moving a window of length *l* over the corpus by one word increments ignoring punctuation, sentence and paragraph boundaries. All words within the window are considered as co-occurring with each other with strengths inversely proportional to the distance between them. After traversing the communication corpus, an accumulated co-occurrence matrix for all the words in a target vocabulary is produced: the semantic space.

More formally, a concept[4] c_i is a vector representation: $c_i = \left\langle w_{c_i p_1}, w_{c_i p_2}, ... w_{c_i p_n} \right\rangle$ where $p_1, p_2, ..., p_n$ are called dimensions of c_i, n is the dimensionality of the HAL space, and $w_{c_i p_i}$ denotes the weight of p_i in vector of c_i. A dimension is termed a property if its weight is greater than zero. A property p_i of a concept c_i is a termed quality property iff $w_{c_i p_i} > \partial$, where ∂ is a non-zero threshold value. Let $QP(c)$ denote the set of quality properties of concept *c*.

3.2 Combining concepts

Concept combination is important as combinations of words in may represent a single underlying concept, for example, *Sun's share price*. An important intuition in concept combination is that one concept can dominate the other. For example, the term "Sun" can be considered to dominate the term "price" because it carries more of the information in the phrase. Given two concepts $c_1 = \left\langle w_{c_1 p_1}, w_{c_1 p_2}, ... w_{c_1 p_n} \right\rangle$ & $c_2 = \left\langle w_{c_2 p_1}, w_{c_2 p_2}, ... w_{c_2 p_n} \right\rangle$, the resulting combined concept is denoted $c_1 \oplus c_2$. The following concept combination heuristic is essentially a restricted form of vector addition whereby quality properties shared by both concepts are emphasized, the weights of the properties in the dominant concept are re-scaled higher, and the resulting vector from the combination heuristic is normalized to smooth out variations due to differing number of contexts the respective concepts appear in.

Step 1: Re-weight c_1 and c_2 in order to assign higher weights to the properties in c_1.

$$w_{c_1 p_i} = \ell_1 + \frac{\ell_1 * w_{c_1 p_i}}{\underset{k}{Max}(w_{c_1 p_k})} \quad \text{and} \quad w_{c_2 p_i} = \ell_2 + \frac{\ell_2 * w_{c_2 p_i}}{\underset{k}{Max}(w_{c_2 p_k})}$$

$$\ell_1, \ \ell_2 \in (0.0, \ 1.0) \text{ and } \ell_1 > \ell_2$$

For example, if $\ell_1 = 0.5$ and $\ell_2 = 0.4$, then property weights of c_1 are transferred to interval [0.5, 1.0] and property weights of c_2 are transferred to interval [0.4, 0.8], thus scaling the dimensions of the dominant concept higher.

Step 2: Strengthen the weights of properties appearing in both c_1 and c_2 via a multiplier α; the resultant highly weighted dimensions constitute significant properties in the resultant combination.

$$\forall (p_i \in QP(c_1) \wedge p_i \in QP(c_2)) \ \mid \ w_{c_1 p_i} = \alpha * w_{c_1 p_i},$$

$$w_{c_2 p_i} = \alpha * w_{c_2 p_i}, \text{ where } \alpha > 1.0$$

Step 3: Compute property weights in the composition $c_1 \oplus c_2$:

$$w_{(c_1 \oplus c_2) p_i} = w_{c_1 p_i} + w_{c_2 p_i}, 1 \le i \le n$$

Step 4: Normalize the vector $c_1 \oplus c_2$. The resultant vector can then be considered as a new concept, which, in turn, can be composed to other concepts by applying the same heuristic.

In order to deploy the information flow model in an experimental setting, the pseudo-queries have to analysed for concept combinations. In particular, the question of which concept dominates which other concept(s) needs to be resolved. As there seems to be no reliable theory to determine dominance, a heuristic approach is taken in which dominance is determined by multiplying the query term frequency (*qtf*) by the inverse document frequency (*idf*) value of the query term. More specifically, query terms can re ranked according to *qtf*idf*. Assume such a ranking of query terms: $q_1,...,q_m$ (m > 1). Terms q_1 and q_2 can be combined using the concept combination heuristic described above resulting in the combined concept $q_1 \oplus q_2$, whereby q_1 dominates q_2 (as it is higher in the ranking). For this combined concept, its degree of dominance is the average of the respective *qtf*idf* scores of q_1 and q_2. The process recurses down the ranking resulting in the composed query "concept" $((..(q_1 \oplus q_2) \oplus q_3) \oplus ...) \oplus q_m)$. This denotes a single vector from which query models can be derived. If there is a single query term (m =1), it's corresponding normalized HAL vector is used for query model derivation.

As it is important to weight pseudo-query terms highly, the weights of query terms which appeared in the initial query were boosted in the resulting query model by adding 1.0 to their score. Due to the way HAL vectors are constructed, it is possible that an initial query term will not be represented in the resulting query model. In such cases,

[4] The term "concept" is used somewhat loosely; it can be envisaged as "term" in the traditional IR sense

the query term was added with a weight of 1.0. Pilot experiments show that the boosting heuristic performs better than the use of only query models without boosting.

3.2 Using the semantic space – information flow

Barwise & Seligman [13] have proposed an account of information flow that provides a theoretical basis for establishing informational inferences between concepts. For example,

share, *price* |- *SUN*

illustrates that the concept "SUN" is carried informationally by the combination of the concepts "share" and "price". Said otherwise, "SUN" *flows* informationally from "share" and "price". Such information flows are determined by an underlying information state space. A HAL vector can be considered to represent the information "state" of a particular concept (or combination of concepts) with respect to a given corpus of text. The degree of information flow between "satellites" and the combination of "space " and "program" is directly related to the degree of inclusion between the respective information states represented by HAL vectors. Total inclusion leads to maximum information flow. Inclusion is a relation ⊆ over the concept space.

Definition 1 (HAL-based information flow)

$$i_{1,} \ldots, i_n \,|\!-\, j \text{ iff degree}(\oplus c_i \subseteq c_j) > \lambda$$

where c_i denotes the conceptual representation of token *i*, and λ is a threshold value. (For ease of exposition, $\oplus c_i$ will be referred to as c_i because combinations of concepts are also concepts).

Note that information flow shows truly inferential character, i.e., concept *j* is not necessarily a dimension of the $\oplus c_i$. The degree of inclusion is computed in terms of the ratio of intersecting quality properties of c_i and c_j to the number of quality properties in the source c_i:

$$\text{degree}(c_i \subseteq c_j) = \frac{\sum_{p_l \in (QP(c_i) \wedge QP(c_j))} w_{c_i p_l}}{\sum_{p_k \in QP(c_i)} w_{c_i p_k}}$$

In terms of the experiments reported below, the set of quality properties $QP_i(c_i)$ in the source HAL vector c_i is defined to be all dimensions with non-zero weight (i.e., ∂ > 0). The set of quality properties $Qj_i(c_j)$ in the target HAL vector c_j is defined to be all dimensions greater than the average dimensional weight within c_j**.** These definitions for determining the quality properties in the source concept c_i and target concept c_j were determined via pilot studies in information flow computation.

2.3 Deriving query models via information flow

Given the pseudo-query $Q=(q_1,...,q_m)$ drawn manually from a standard S in the prescriptive model P, a query model can be derived from Q in the following way:

- Compute degree($\oplus c_i \subseteq c_t$) for every term *t* in the vocabulary, where $\oplus c_i$ represents the conceptual combination of the HAL vectors of the individual query terms $q_i, 1 \le i \le m$ and c_t represents the HAL vector for term *t*.
- The query model $Q' = \langle t_1 : f_1, \ldots, t_k : f_k \rangle$ comprises the top *k* information flows

Observe that the weight f_i associated with the term t_i in the query model is not probabilistically motivated, but denotes the degree to which we can infer t_i from *Q* in terms of underlying HAL space.

4. BLOG DATA

Blog data, as input to computational analysis as distinct from human comprehension, is inherently "dirty": it can consist of anything from a URL, presumably as aid to the memory of the author and often with a longer title explaining something, or it can be a long-winded polemic in the first person. Nardi et al [17] found that people blog for (at least) five reasons – documenting one's life, providing commentary and opinions, expressing deeply felt emotions, articulating ideas through writing, and forming and maintaining community forums. While humans find it relatively easy to navigate the morass, find interesting elements and determine the worth of data, comparatively this is almost impossible for current computer systems.

A vital element is a filter to identify "interesting" blog entries which would be used to populate the semantic space(s). "Interesting" is determined by the particular person doing the searching, or the particular problem. For example, if the question is one of compliance—is a particular blog entry compliant with Sun's policies—the filter would provide very different entries than if an individual were interested in a particular Sun product.

It is feasible to produce filters which could identify the five+ (non-exclusive) motivations as only some of these are relevant to policy conformance checking. It is also important to filter the difference between a wilful breaking of policy and an inadvertent one.

Many such situational-based filters are possible. The focus of these experiments was to apply one such filter to the blog entries. Note that for checking of blog entry compliance, the filter may be enacted *prior* to blog entry publication (as in Figure 1) or afterwards. While the method we describe could be used in both ways, we envisage that human invigilators would prefer to peruse candidate entries at certain times during the day rather than being interrupted for each possibility. This is of course offset by the desire to preserve the currency of the entries.

4.1 Experimental data

We examined all entries from the Sun blog RSS feed from 19 July to 9 August (22 days) 2004. There were 1507 RSS entries at an average of 68.5 per day (2.8 per hour); the minimum was 17 entries on July 31st. However, on two days-the 26th and 27th of July-there were 404 and 140 entries respectively. This was due to discussion about a new product about to be released ([16] have some further insights into these phenomena). Figure 1 charts the entries over time.

The size of the vocabulary (stop words removed) was 24,841 words. As we only examined entries from the RSS feed, we were not able to account for comments submitted to existing blog entries, and other associated text that did not appear in the RSS. Where available, this could augment the analysis.

It is important to note that no set of disconforming data was provided. We do not have details of any blog entries that were filtered prior to publication, and cannot guarantee that those that we worked with are all still extant. All experimental work was conducted on blog entries that were publicly available at the time. We do not know if Sun would consider any particular entry we have discussed disconformant. In this way, although our analysis lets us work unfettered by internal prejudice, we may miss nuances that an internal assessor would not.

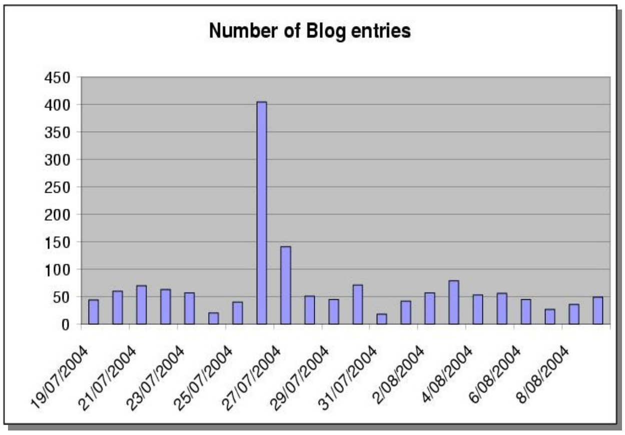

Figure 2: Number of blog entries over experiment time

5. ANALYSIS

To provide a flavour of the data in the form of semantic spaces, two tables show the results of computations creating semantic spaces over the entire collection: table 1 shows the words with the largest number of dimensions (ie. words used in many contexts); table 2 shows the "largest" explicit dimensions of the "sun" vector.

Table 1: "Sun" vector top dimensions

sun	1008	back	387
solaris	662	entry	319
new	651	things	314
java	651	great	313
open	556	dtrace	300
good	516	software	295
work	474	blog	277
people	461	code	266
system	420	linux	265
don	415	...	...
source	397		

Table 2: Top "sun" vector dimensions (cols 1-2) and nearest concepts (cosine; cols 3-4); 4511 dimensions, $\bar{x}$: 19.2, σ^2: 54.6

sun	2266.00	java	0.84
java	1443.00	working	0.75
open	819.00	microsystems	0.72
solaris	817.00	workstation	0.72
system	565.00	product	0.72
source	529.00	work	0.71
new	459.00	lot	0.71
work	456.00	community	0.70
working	438.00	customers	0.70
people	385.00	system	0.70
company	379.00	product	0.70
customers	306.00	people	0.69
server	299.00	developer	0.69
desktop	292.00	company	0.69
blog	278.00	ibm	0.65
software	277.00	desktop	0.65
good	275.00	software	0.65
product	275.00	employees	0.65
ray	273.00	developers	0.64
microsystems	266.00	worked	0.64
lot	262.00	new	0.64
community	252.00	reason	0.64
linux	249.00	part	0.64
cluster	243.00	ray	0.63
employees	240.00	vendor	0.63
things	232.00	customer	0.63
products	225.00	cluster	0.62
business	223.00	hardware	0.61
support	219.00	new	0.61
ibm	213.00	don	0.61
workstation	213.00	companies	0.61
...	...	...	...

5.1 Information flow based query expansion

The "Financial rules" section in the policy (Appendix A) states:

> *There are all sorts of laws about what we can and can't talk about. Talking about revenue, future product ship dates, road maps, or our share price is apt to get you, or the company, or both, into legal trouble.*

The challenge is to mimic human's ability to interpret the above standard while considering a certain blog entry.

A semantic space H_B can be constructed from the blog corpus B using the Hyperspace Analogue to Language model. The goal is to provide a semantic representation $\sigma(c)$ for concept c which will be used as a "query" to match incoming blog entries. If the match score is above a certain threshold it can be flagged for human perusal.

In our previous work, encouraging improvements in retrieval precision were produced by information flow based query expansion [2]. For the purposes of illustration, we focus on the financial area by characterizing it with the concept "share price", which is a noun phrase. Our concept combination heuristic produces a semantic representation of the compound using the individual semantic representations σ(share) and σ(price). (See [15] for more details of this heuristic). Each of the two pseudo-queries was expanded using information flow. Table 3 shows the top information flows from the concept "share price". The top 65 information flows (empirically determined) were used to expand the pseudo-query. The resulting expanded query was matched against blog entries which were ranked on decreasing order of retrieval status score. In order to facilitate matching each blog was indexed using the BM-25 term weighting score[5], with stop words removed. Both query and document vectors were normalized to unit length. Matching was realized by the dot product of the respective vectors and the top five ranked blog entries were chosen. This threshold was chosen as we assume that human judges will not want to manually peruse rankings much longer than this.

Table 3: Information flows from the concept "share price"

Flow	*Value*
price	0.77
share	0.68
sun	0.59
good	0.51
back	0.44
don	0.44
software	0.53
...	...

5.2 Experimental Results

Due to the small number of pseudo-queries it is not warranted to present a precision-recall analysis. A much larger experimental setting would be required.

Discussion will proceed based on anecdotal evidence. The following document (next column) was ranked second with respect to the pseudo-query "share price" and is the most interesting of the top five.

5.3 Discussion

The retrieval of the above document demonstrates the potential of information flow query expansion. Note how the phrase "share price" does *not* appear in this blog entry, but is clearly about a strongly related concept (stock option). This example also shows how information flow based query expansion facilitates the promotion of potentially disconformant blogs in the retrieval ranking when there is little or no term overlap between the pseudo-query and blog entry. In order to place this claim in perspective, we expanded the pseudo-query "share price" with a highly respected probabilistic retrieval model - the BM25 model [14], and a query expansion technique - the Robertson's Term Selection Value (TSV) [14]. Both techniques were unable to rank the above document in the top five.

<CONTEXT ID="//blogs.sun.com/roller/page/pdiamond/20040624#stock options why not expense">
</CONTEXT>

Stock Options - Why not expense them?

Just came from the rally in Palo Alto to oppose FASB ruling that stock options should be expensed. For those who do NOT have access to stock options, the answer seems pretty simple:

> "These people are making lots of money off stock options, taking advantage of opportunities we don't have and inaccurately reflecting their expense on their companies' bottom lines. Of course they should be counted as an expense when they are granted"

I'm sure a lot of this is also reflective of the abuses which have been widely reported, of CxOs making million$ while their companies went down the tubes.

Now here's another view of reality - for those of us who have

- made some money (thank you, Netscape) and
- not made any yet (I am still optimistic, Sun),

it also seems pretty obvious.

All those options we have been granted which we do NOT exercise, because they are "underwater", e.g.:

- Netscape /AOL options at $75 when the stock price was $20,
- current Sun options at $12 (and I know many people with options well above that price) with the stock a little over $4, are irrelevant to anyone. They are no more expense to the companies which granted them than they are profit to the employees who are not exercising them.

If and when they are exercised, then let's talk about how the companies should expense the benefit received by the employees. I admit to being ignorant as to how this is handled today. This seems to be a much more relevant issue than trying to assess some current value on some theoretical future benefit, which in many cases will either not happen, or will occur at a totally unpredictable level.

June 24, 2004 04:07 PM PDT Permalink

5.4 Temporal topics of Pseudo-queries

Tracking the temporal profile of a pseudo-query over time can help visualize blog activity around a topic relevant to detecting disconformance. Figure 3 depicts the probability of the pseudo-query "share price" over time. The underlying theory combines information flow based query expansion [2] with document language models. The probabilities of queries were calculated from top ten documents retrieved by the information flow model and then smoothed using a back-off model based on collection statistics. The spikes in the figure depict localized

[5] BM-25 represents state-of-the-art in term weighting, e.g. [14]

probabilities of the topic which can be used to localize activity around a pseudo-query. Such localities may warrant closer inspection for disconformance.

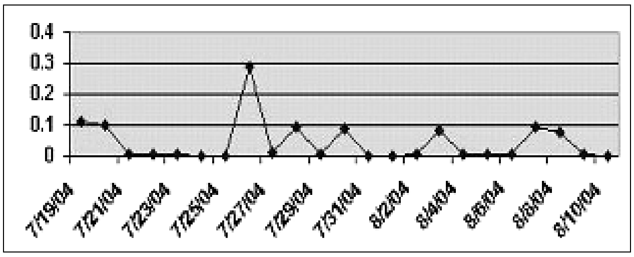

Figure 3: Temporal profile for topic "share price"

5.5 Optimal projections

The approach here is to assume that blogs disconforming to a standard S_i will cluster around a given axis, or somehow project differently into the semantic space than conforming blog entries. Dimensional reduction approaches may gain some purchase, for example independent component analysis or projection pursuit. Further investigation is required.

4. CONCLUSION

This article deals with the problem of providing automated support for the detection of disconformant blog entries with respect to a publishing policy. The problem is considered from a normative perspective. The detection of disconformant blog entries has an abductive character. Automated support for detecting disconformant blogs is realized via query expansion, the goal of which is to abduce salient terms in relation to pseudo-query representations of publishing standards. The expanded pseudo-queries are computed vie information flows through a high dimensional semantic space derived from the blog corpus.

Anecdotal evidence suggests that information flow based query expansion may be promising in regard to retrieving disconformant blog entries, which can then be manually examined for a final judgment. The case study reported in this paper suggests that the problem of furnishing (semi-) automated support for the detection of disconformat blog entries to be a challenging one requiring further investigation using non-supervised approaches.

ACKNOWLEDGMENTS

The work reported in this paper has been funded in part by the Co-operative Research Centre for Enterprise Distributed Systems Technology (DSTC) through the Australian Federal Government's CRC Programme (Department of Education, Science, and Training). The research was funded by a Discovery grant from the Australian Research Council. Thanks to the members of the Research Management Team and participants.

REFERENCES

[1] Gabbay, D. and Woods, J. (2003): Normative models of rational agency: the theoretical disutility of certain approaches. *Logic Journal of the IGPL.* 11:597-613

[2] Bruza, P.D and Song, D. (2002): Inferring query models by computing information flow. In *Proceedings of the 11th International Conference on Information and Knowledge Management (CIKM 2002)* ACM Press, pp.260-269.

[3] Nonaka, I. and Takeuchi, H. (1995): *The Knowledge-Creating Company*, OUP, New York

[4] Freyd, J. (1983). Shareability: the social psychology of epistemology. *Cognitive Science*.**7**:191-210

[5] Gärdenfors, P. (2000): *Conceptual Spaces: the Geometry of Thought.* MIT Press, London, 2000

[6] Ducheneaut, N. and Bellotti, V. (2003). 'Ceci n'est pas un objet? Talking about things in email. *Journal of Human-Computer Interaction (special issue)* **18**(1-2): 85-110.

[7] McArthur, R. and P. Bruza (2003). Dimensional Representations of Knowledge in Online Community. In *Chance Discovery*. Y.Ohsawa & P.McBurney, Springer**:** 98-112

[8] McArthur, R. and P. Bruza (2003). Discovery of Tacit Knowledge and Topical Ebbs and Flows within the Utterances of Online Community. In *Chance Discovery.* Y. Ohsawa and P. McBurney, Springer**:** 115-131.

[9] McArthur, R. and P. Bruza (2003). Discovery of Implicit and Explicit Connections between People using Email Utterance. In *Eighth European Conference on Computer-Supported Cooperative Work (ECSCW)*, Helsinki, Finland, Kluwer.

[10] Burgess, C., Livesay, K. and Lund, K. (1998): Explorations in context space: words, sentences, discourse. *Discourse Processes*, v25, pp.211-257

[11] Burgess, C. and K. Lund (1997b). Representing Abstract Words and Emotional Connotation in a High-Dimensional Memory Space. *Cognitive Science*.

[12] Lund, K., C. Burgess and R. A. Atchley (1995). Semantic and Associative Priming in High-Dimensional Semantic Space. *Cognitive Science*, Erlbaum Publishers, Hillsdale, N.J.

[13] Barwise, J. and Seligman, J. (1997) *Information Flow*. Cambridge University Press.

[14] Robertson, S., Walker, S., Jones, S., Hancock-Beaulieu, M. and Gatford, M. (1995) Okapi at TREC-3. In *Proceedings of TREC-3* 1995. Available at trec.nist.gov.

[15] Song,D. and Bruza, P. (2003) Towards context sensitive information inference. *Journal of the American Society for Information Science and Technology*, 54(3):321-334.

[16] Kumar, R., Novak, J., Raghavan, P. and Tomkins, A. (2004): Structure and evolution of blogspace. *Communications of the ACM*, 47(12) pp35-39

[17] Nardi, B., Shiano, D., Gumbrecht, M. and Swartz, L. (2004) Why we blog? *Communications of the ACM.* v47(12)

APPENDIX A: SUN'S BLOGGING POLICY

Advice By speaking directly to the world, without benefit of management approval, we are accepting higher risks in the interest of higher rewards. We don't want to micro-manage, but here is some advice.

It's a Two-Way Street The real goal isn't to get everyone at Sun blogging, it's to become part of the industry conversation. So, whether or not you're going to write, and especially if you are, look around and do some reading, so you learn where the conversation is and what people are saying.

If you start writing, remember the Web is all about links; when you see something interesting and relevant, link to it; you'll be doing your readers a service, and you'll also generate links back to you; a win-win.

Don't Tell Secrets Common sense at work here; it's perfectly OK to talk about your work and have a dialog with the community, but it's not OK to publish the recipe for one of our secret sauces. There's an official policy on protecting Sun's proprietary and confidential information, but there are still going to be judgment calls.

If the judgment call is tough—on secrets or one of the other issues discussed here—it's never a bad idea to get management sign-off before you publish.

Be Interesting Writing is hard work. There's no point doing it if people don't read it. Fortunately, if you're writing about a product that a lot of people are using, or are waiting for, and you know what you're talking about, you're probably going to be interesting. And because of the magic of hyperlinking and the Web, if you're interesting, you're going to be popular, at least among the people who understand your specialty.

Another way to be interesting is to expose your personality; almost all of the successful bloggers write about themselves, about families or movies or books or games; or they post pictures. People like to know what kind of a person is writing what they're reading.

Once again, balance is called for; a blog is a public place and you should try to avoid embarrassing your readers or the company.

Write What You Know The best way to be interesting, stay out of trouble, and have fun is to write about what you know. If you have a deep understanding of some chunk of Solaris or a hot JSR, it's hard to get into too much trouble, or be boring, talking about the issues and challenges around that.

On the other hand, a Solaris architect who publishes rants on marketing strategy, or whether Java should be open-sourced, has a good chance of being embarrassed by a real expert, or of being boring.

Financial Rules There are all sorts of laws about what we can and can't say, business-wise. Talking about revenue, future product ship dates, roadmaps, or our share price is apt to get you, or the company, or both, into legal trouble.

Quality Matters Use a spell-checker. If you're not design-oriented, ask someone who is whether your blog looks decent, and take their advice on how to improve it.

You don't have to be a great or even a good writer to succeed at this, but you do have to make an effort to be clear, complete, and concise. Of course, "complete" and "concise" are to some degree in conflict; that's just the way life is. There are very few first drafts that can't be shortened, and usually improved in the process.

Think About Consequences The worst thing that can happen is that a Sun sales pro is in a meeting with a hot prospect, and someone on the customer's side pulls out a print-out of your blog and says "This person at Sun says that product sucks."

In general, "XXX sucks" is not only risky but unsubtle. Saying "Netbeans needs to have an easier learning curve for the first-time user" is fine; saying "Visual Development Environments for Java suck" is just amateurish.

Once again, it's all about judgment: using your weblog to trash or embarrass the company, our customers, or your co-workers, is not only dangerous but stupid.

Disclaimers Many bloggers put a disclaimer on their front page saying who they work for, but that they're not speaking officially. This is good practice, but don't count it to avoid trouble; it may not have much legal effect.