Enabling Management Oversight in Corporate Blog Space

[1]Dawei Song [2]Peter Bruza [3]Robert McArthur [4]Tim Mansfield

[1] Knowledge Media Institute, The Open University
Walton Hall, Milton Keynes, MK7 6AA, United Kingdom
d.song@open.ac.uk

[2] School of Information Systems, Queensland University of Technology
GPO Box 2434, Brisbane, QLD 4001, Australia
p.bruza@qut.edu.au

[3] CSIRO ICT Centre
Building 108, Australian National University Campus North Road, Acton ACT 0200, Australia
Robert.McArthur@csiro.au

[4] Distributed Systems Technology Centre
The University of Queensland, QLD 4072 Australia
timbomb@timbomb.net

Abstract

When a modern corporation empowers its staff to use blogs to communicate with colleagues, partners, suppliers and customers, the role of management in exercising oversight and guidance over the public speech of staff becomes dramatically challenged. This paper describes a computational solution to the interpretation of human-readable blog publishing policy documents into semi-automatic disconformance checking of corporate blog entries. The disconformance interpretation is regarded as an abductive reasoning, which is operationalized by information flow computations. Using a socio-cognitively motivated representation of shared knowledge, and applying an appropriate information flow inference mechanism 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 judgment on whether they should be published. Experiments on data from a public corporate blog demonstrate an encouraging performance of the proposed methodology.

1. Introduction

Managers of organisations control what is the official word of the body versus what an employee personally has presented. When a modern corporation empowers its staff to use blogs and emails to communicate with colleagues, partners, suppliers or customers, the traditional role of management in exercising oversight and guidance over the public speech of staff becomes dramatically challenged. There has been an increasing drive for knowledge sharing in order to boost the knowledge capital and performance of organizations. Blogs have emerged strongly as a technology to support this drive. However, there is a trade off, the open-ness required for knowledge sharing impinges on management's ability to monitor and control what is being published in employee blogs. A particular example is the Sun Microsystem's need to issue a policy stating confidential matters, or opinions about financial issues relevant to SUN should not be blogged. 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)*." Due to limit of space, we do not reproduce it here, but just show the so-called financial issues:

> *"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." [Bray 2004]*

[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)

It is obvious that for the management trying to review thousands of blog entries a week to ensure that they do not expose confidential information or break financial rules consumes an unacceptable amount of time.

This paper aims for developing techniques to locate a “review set” of electronic communication entries which seem likely to infringe a given policy on public discourse. We propose a computational solution to the interpretation of human-readable policy documents into automatic disconformance checking. The concepts relevant to the policy-breaking terms, for example, “revenue”, “future product ship dates”, “road maps” and “share price”, will be used as “pseudo-queries” to match blog entries. If the match score is above a certain threshold it can be flagged for human perusal. The challenge is to mimic human's ability to interpret the above standard while considering a certain blog entry. For example, a blog entry related to “share price” may not explicitly contain either of the terms “share” and “price”, but instead contain “stock” or “options”. The challenge is accessing the background knowledge around a pseudo-query and inferring implicitly related terms. We refer to such inference as “information inference” [Song and Bruza 2003]. By using a socio-cognitively motivated knowledge representation computed form the blog space, together with information inference, a policy checking mechanism is detailed to automatically detect potentially disconforming blog entries. Candidate non-conforming blog entries are flagged for a human to make a judgment on whether they should be published. The benefits are a significant lessening of work for humans to evaluate each blog entry. Instead, only a subset (referred to as review set in this paper) is required to be vetted by a person. Figure 1 shows the overall architecture of our approach.

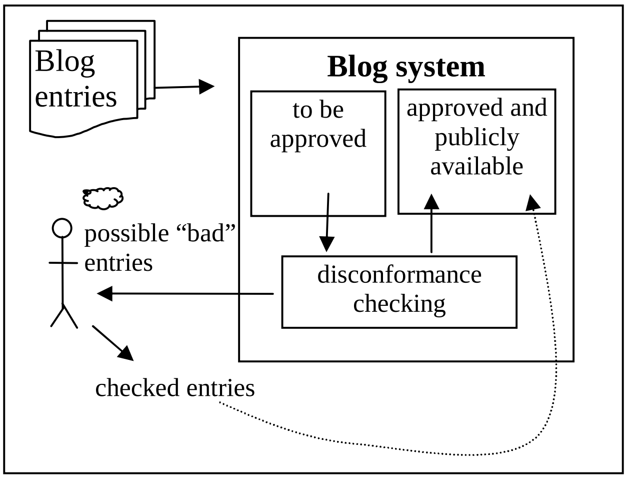

Figure 1: Architecture

The remainder of this paper is organized as follows. Section 2 starts with the notion of normative disconformance and describes how the disconformance can be interpreted in form of abduction. Section 3 then discusses how this can be operationalized by computing information flows from the disconforming concepts based on a vector representation of knowledge via semantic spaces. By treating the policy-breaking terms as pseudo-queries, the disconformance checking can be implemented as automatic query expansion based on computations of information flow through the semantic space. Experimental results of examining Sun’s blog data with respect to financial disconformance are reported in Section 4, followed by a discussion in Section 5. Section 6 concludes the paper.

2. Disconformance Detection as Abductive Reasoning

Let N be a normative model comprising principles (or standards) S_1, ..., S_n. Let B be a piece of augmentative behaviour. Let B disconform with principle S_i. If S_i is genuinely normative then B is a mistake (at a minimum) [Gabbay and Woods 2003].

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.

People write blog entries to communicate. In all communication, there are both explicit and tacit parts to the message. 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 mentioning 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).

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 very promising, particularly for short queries [Bruza and Song 2002].

3. Operational Disconformance Detetection

Previous work [McArthur and Bruza 2003] 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, Hyperspace Analogue to Language (HAL) below. We will then discuss ways of using the semantic space in the context of compliance and blog data.

3.1 Knowledge Representation Via HAL

Hyperspace Analogue to Language (HAL) is a model and technique to populate a semantic space [Burgess et al. 1998; Burgess and Lund 1997; Lund et al. 1995]. HAL produces vectorial representations of words in a high dimensional space that seem to correlate with the equivalent human representations. For example, word associations computed on the basis of HAL vectors seem to correlate well with human word association judgments.

What HAL does is to generate a word-by-word co-occurrence matrix from a large text corpus via a *L*-sized sliding window: All the words occurring within the window are considered as co-occurring with each other. By moving the window across the text corpus, an accumulated co-occurrence matrix for all the words in a certain vocabulary is produced. The strength of association between two words is inversely proportional to their distance. Given two words, whose distance within the window is *d*, the weight of association between them is computed by ($L – d + 1$). After traversing the corpus, an accumulated co-occurrence matrix for all the words in a target vocabulary is produced. HAL is direction sensitive: the co-occurrence information for words preceding every word and co-occurrence information for words following it are recorded separately by its row and column vectors.

The quality of HAL vectors is influenced by the window size; the longer the window, the higher the chance of representing spurious associations between terms. A window size of eight or ten has been used in various studies [Burgess et al. 1998, Bruza and Song 2002, Song and Bruza 2001]. Accordingly, a window size of 10 will also be used in the experiments reported in this paper.

More formally, a concept[4] *c* is a vector representation: $c = < w_{cp_1}, w_{cp_2}, ..., w_{cp_n} >$ where $p_1, p_2, ..., p_n$ are called dimensions of *c*, *n* is the dimensionality of the HAL space, and w_{cp_i} denotes the weight of p_i in the vector representation of *c*. In addition, it is useful to identify the so-called *quality properties* of a HAL-vector. Intuitively, the quality properties of a concept or term *c* are those terms which often appear in the same context as *c*. Quality properties are identified as those dimensions in the HAL vector for *c* which are above a certain threshold (e.g., above the average weight within that vector). A dimension is termed a property if its weight is greater than zero. A property p_i of a concept c is a termed a *quality property* iff $w_{cp_i} > \partial$, where ∂ is a non-zero threshold value. From a large corpus, the vector derived may contain much noise. In order to reduce noise, in many cases only certain quality properties are kept. Let $QP_\partial(c)$ denote the set of quality properties of concept *c*. $QP_\mu(c)$ will be used to denote the set of quality properties above mean value, and $QP(c)$ is short for $QP_0(c)$.

Different words can be combined to form more complex concepts like "share price". A vector is also obtained for this latter by combining the HAL vectors of the individual terms. A simple method is to add the vectors of the terms. In this article, however, we employ a more sophisticated concept combination heuristic [Bruza and Song 2002]. It can be envisaged as a weighted addition of underlying vectors modulo the intuition that in a given concept combination, some terms are more dominant than others. For example, the combination "share price" is more "share-*ish"* than "price-*ish*". Dominance is determined by the specificity of the term.

[4] The term "concept" is used somewhat loosely to emphasize that a HAL space is a primitive realization of a conceptual space [Gärdenfors 2000]

In order to deploy the concept combination in an experimental setting, dominance is determined by its inverse document frequency (*idf*) value of the term. More specifically, terms can re-ranked according to its *idf*. Assume such a ranking of terms: $t_1, \ldots, t_{m.}$ (m > 1). Terms t_1 and t_2 can be combined using the concept combination heuristic resulting in the combined concept $t_1 \oplus t_2$, whereby t_1 dominates t_2 (as it is higher in the ranking). For this combined concept, its degree of dominance is the average of the respective *idf* scores of t_1 and t_2. The process recurses down the ranking resulting in the composed "concept" $((..(t_1 \oplus t_2) \oplus t_3) \oplus \ldots) \oplus t_m)$. If there is a single term (m =1), it's corresponding normalized HAL vector is used for combination vector.

We will not give a more detailed description of the concept combination heuristic, which can be found in [Bruza and Song 2002]. Its intuition is summarized as follows:

- Quality properties shared by both concepts are emphasized,
- The weights of the properties in the dominant concept are re-scaled higher,
- The resulting vector from the combination heuristic is normalized to smooth out variations due to differing number of contexts the respective concepts appear in.

3.2 HAL-based Information Flow

Information inference refers to how certain concepts, or combinations of concepts, carry information about another concept. For example,

share, price |- *stock*

illustrates that the concept "stock" is carried informationally by the combination of the concepts "share" and "price". A HAL vector can be considered to represent the information "state" [Barwise and Seligman 1997] of a particular concept or combination of concepts with respect to a given corpus of text. The degree of information flow is directly related to the degree of inclusion between the respective information states represented by HAL vectors. Total inclusion leads to maximum information flow.

More formally, let $i_1, \ldots, i_k (k > 0)$ represent source concepts. What is the information flow from the combination of these concepts to an arbitrary concept *j*? Let $c_{i_1} \oplus \ldots \oplus c_{i_k}$ represent the vector combining the respective vector representations of concepts i_1 to i_k. For ease of exposition, the resulting vector will referred to as c_i because combinations of concepts are also concepts. Let c_j be the vector representation corresponding to the target concept *j*. 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.

The underlying idea of this definition is to make sure that a majority of the most important quality properties of c_i appear in c_j (Song and Bruza 2003).

$$\text{degree}(c_i, c_j) = \frac{\sum_{p_l \in (QP_\mu(c_i) \cap QP(c_j))} w_{c_i p_l}}{\sum_{p_k \in QP_\mu(c_i)} w_{c_i p_k}}$$

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 $QP_j(c_j)$ in the target HAL vector c_j is defined to be all dimensions greater than the average dimensional weight within c_j.

3.3 Deriving Pseudo-query Models via Information Flow

Given the pseudo-query Q=(q_1,...,q_k) 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(c_i, c_t) for every term t in the vocabulary, where c_i represents the conceptual combination of the HAL vectors of the individual query terms q_i, $1 \le i \le k$ and c_t represents the HAL vector for an arbitrary term t.
- The query model $Q' = \langle t_1 : f_1, \ldots, t_m : f_m \rangle$ comprises m terms with the highest information flow from the query Q, where f_i represents the degree of the flow.

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. Experiments

Blog entries can consist of anything from a URL, presumably as aid to the memory of the author and often with a longer title explaining something, to a long-winded polemic in the first person [Nardi et al 2004]. As a result, blog data, as input to computational analysis as distinct from human comprehension, is inherently "dirty".

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.

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 Methodology

In our experiment, we take a pseudo-relevance feedback approach. The test query is first matched against the blog entries using a traditional probabilistic information retrieval model, namely BM-25 [Robertson et al. 1995]. The top N ranked documents are considered as relevant and then used to build a semantic space, which is much more precise and compact than the one built from the entire corpus.

In our previous work, encouraging improvements in retrieval precision were produced by information flow based query expansion [Bruza and Song 2002; Song and Bruza 2003]. For the purposes of illustration, we focus on the financial area by characterizing it with the concepts “revenue”, “future product ship dates”, “road maps” and “share price”, which form four pseudo queries. Our concept combination heuristic produces a semantic representation of the compounds using the individual semantic representations, e.g., for “share” and “price”. Each of the pseudo-queries can be expanded using information flow. The top 65[5] information flows 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, 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 100 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.

4.2 Test Data and Method

In late May 2005, we used the planetsun.org RSS feed to download 2500 blog entries over a period of a couple of weeks. It is important to note that 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 pre-processed the textual data to remove XML and HTML tags and punctuation and formatted it into the input format for the experimental tool.

Based on some additional research, we then prepared four test entries which we believed did not conform to the rules. Each entry was written with the four forbidden topics (revenue, future product ship dates, road maps, Sun’s share price) as its theme. We showed these to Tim Bray and Simon Phipps, who are SUN’s blog policy makers. They confirmed that each test entry broke the rules. These four were then pre-processed and added to the corpus.

The corpus was recoded with numeric identifiers so that it was difficult for the experimenter to identify which were the test entries. The test entries are numbered as 2523, 2524, 2525, and 2526. The full details of test entries are listed in the Appendix. A member of the team not involved in the experiment kept the code table “in escrow”, so that the tests were blind. The entire corpus was handed to the experimenter to try to uncover the non-compliant entries.

The experimenter conducted five runs of the technique described in Section 3 with different numbers (N) of documents in the relevance set used for feedback. A state-of-the-art information retrieval system based on BM-25 was used for comparison. The base line system used Robertson’s Term Selection Value (TSV) to drive query expansion [Robertson et al. 1995]. It should be noted that the baseline system used in the experiments is representative of a system that has consistently performed well within information retrieval research.

4.3. Experimental Results

The following tables summarize our experimental results by showing the rankings of the test entries in the top 100 matching documents for each pseudo-query.

Rank	Pseudo-query	Test Entry
7	“revenue”	2526
44	“future product ship date”	2526
56	“roadmap”	2524

Table 1: Results for Information Flow (N10)

Rank	Pseudo-query	Test Entry
7	“revenue”	2526
60	“future product ship date”	2526
56	“roadmap”	2524
65	“share price”	2523

[5] This setting produced the best results in our previous experiments [Bruza and Song 2002].

Table 2: Results for Information Flow (N20)

Rank	Pseudo-query	Test Entry
7	“revenue”	2526
40	“future product ship date”	2526
71	“future product ship date”	2525
56	“roadmap”	2524
65	“share price”	2523

Table 3: Results for Information Flow (N30)

Rank	Pseudo-query	Test Entry
7	“revenue”	2526
61	“future product ship date”	2526
67	“future product ship date”	2525
56	“roadmap”	2524
65	“share price”	2523

Table 4: Results for Information Flow (N40)

Rank	Pseudo-query	Test Entry
7	“revenue”	2526
47	“future product ship date”	2523
90	“future product ship date”	2526
56	“roadmap”	2524
60	“share price”	2523

Table 5: Results for Information Flow (N50)

Rank	Pseudo-query	Test Entry
53	“future product ship date”	2523
89	“future product ship date”	2526

Table 6: Results for BM-25

Due to the small number of pseudo-queries it is not warranted to present a precision-recall analysis.

5. Discussion

The challenge is to mimic the human’s ability to interpret the standards listed in the “Financial rules” (Section 1) while considering a certain blog entry.

Tables 2 to 5 show that the detection of disconformant blogs by information flow pseudo-query expansion seems fairly stable as a function of N, the number of blog entries used to construct the underlying semantic space. The experiments, though small, seem to suggest that somewhere around forty blog entries suffice to produce a reliable semantic space from which pseudo-query expansion terms can be computed. It is important to note that the underlying retrieval engine is an issue here. In our case, we employed a state-of-the-art IR engine with good average precision. A lesser IR engine may necessitate larger values of N. More extensive experiments will need to bear this out.

Tables 4 shows the best performance of pseudo-query that all four disconformant blog entries were detected within the top 70 ranked entries after the blogs were re-ranked by the expanded pseudo-query by information flow. A state of the art retrieval system based on BM-25 term weighting and query expansion only detected two of the four disconformant blogs within the top 100 ranked entries.

These results indicate that it seems to be possible to effectively reduce the compliance-checking task for a sample size of 2500 entries, which in Sun’s case represented a couple of weeks of blog entries, down to a review set of only a couple of hundred entries.

Encouraging results have been produced by information flow based pseudo-query expansion. 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. (See [Bruza and Song 2002]). Each of the two pseudo-queries was expanded using information flow. Table 7 shows the top information flows from the concept “share price”.

Flow	*Value*
price	0.88
share	0.79
sun	0.67
good	0.60
…	…
stock	0.52
software	0.50
...	...

Table 7: Information flows from the concept “share price"

The following document is one of the test entries which were retrieved with respect to the pseudo-query “share price”.

> ***Entry #2523***
>
> At a party last night...
>
> ... various people tried to engage me in conversations about the "lamentable state" of Sun's **stock.** These people who remember back in 2000 when it was bumping around near the $50 mark. I tried to explain that that was back in the middle of the bubble and probably artificially inflated, but to little effect. I really wanted to leak stuff that I've heard which all points to us getting back up near $10 in the very near future... but I can't.
>
> Time for lunch.

The retrieval of the above blog entry demonstrates the potential of information flow query expansion. Note how the phrase “share price” does *not* appear in this 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, the expanded pseudo-query “share price” using the baseline system was unable to rank the disconformant blog within the top 100 retrieved entries.

Information flow computations are dependent on the underlying HAL space, which is in turn dependent on the collection it is built upon. It should reflect the context in which the disconformant blogs are checked. As the HAL space has an additive property, it can always be built and enriched accumulatively upon incoming blog entries.

On the other hand, we observed that some non-disconformant entries are misidentified in the review set. This is inherent to the underlying matching function which measures the overlap of terms between the blog entries and an expanded pseudo-query. The latter is often an approximate interpretation of the corresponding policy. The more accurate the interpretation is, the higher the truly disconformant entries are ranked. Therefore, our future work will be focused on developing more effective disconformance interpretation mechanisms.

6. Conclusions

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 blog publishing policy. The expanded pseudo-queries are computed via information flow through a high dimensional semantic space computed from the blog corpus.

Experimental results suggest that information flow based query expansion is promising in regard to retrieving disconformant blog entries, which can then be manually examined for a final judgment. Further work may reduce the safe size of the review set even further so that effective oversight of enterprise-scale blogging could remain feasible as more staff use blogs as a means of communication.

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 un-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

Barwise, J. and Seligman, J. (1997) *Information Flow: The Logic of Distributed Systems.* Cambridge University Press.

Bray, T. et al, (2004) Sun Policy on Public Discourse, http://www.sun.com/aboutsun/media/blogs/policy.html

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.

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

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

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

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

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

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.

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

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.

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

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.

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.

Appendix – Test Entries

Entry #2523

At a party last night...

... various people tried to engage me in conversations about the "lamentable state" of Sun's stock. These people who remember back in 2000 when it was bumping around near the $50 mark. I tried to explain that that was back in the middle of the bubble and probably artificially inflated, but to little effect. I really wanted to leak stuff that I've heard which all points to us getting back up near $10 in the very near future... but I can't.

Time for lunch.

Entry #2524

Watch This Space

A lot of us at Sun get really tired watching the commentary over at <a href="http://www.slashdot.org">Slashdot</a>. They don't understand out our various open source efforts, they just don't get Java... but the thing that really gets me is when they try to write us off based on quarterly revenues! Bunch of 17 year-old geeks who don't get market ebbs and flows.

I mean, I know last quarter was a little shy of the $2.7 billion everyone was anticipating, but I've seen the projections for next quarter. Based on all the stuff we've got out the door in the last year (DTrace!), we will hit $3billion next quarter. That'll mean a healthy profit on top of our growing rep in community leadership.

Sun rocks!

Entry #2525

We're joining Eclipse!

I just overheard a conversation, so I thought I'd break some news that a lot of people have been hoping for. Sun is joining Eclipse.org and offering full support for the Eclipse platform. It will be officially announced at the end of the month.

The sad part of that news is that NetBeans 4.1 will be the final release of NetBeans that Sun will support. What will happen to NetBeans now is anyone's guess... I guess.

Entry #2526

Why Does Anyone Care About Longhorn?

I totally don't get why anyone pays attention to Microsoft playing catch-up and claiming innovation in the same breath. The buzz around WinFS is particularly weird.

The kind of stuff that we're doing on Honeycomb is so innovative that it just leaves Redmond in the dust! Rich metadata, failure impregnability, it's really cool. Add to that the fact that we'll have the basic product released as a beta in September this year and full release in January 2006 - months ahead of Longhorn and I really wonder why we're not getting all that press!