Buzz Monitoring in Word Space

Magnus Sahlgren[1,2] and Jussi Karlgren[1,2]

[1] Gavagai AB
[2] SICS (Swedish Institute of Computer Science)
Box 1263, SE-16429 Kista, Sweden
{mange,jussi}@gavagai.se

Abstract. This paper discusses the task of tracking mentions of some topically interesting textual entity from a continuously and dynamically changing flow of text, such as a news feed, the output from an Internet crawler or a similar text source — a task sometimes referred to as *buzz monitoring*. Standard approaches from the field of information access for identifying salient textual entities are reviewed, and it is argued that the dynamics of buzz monitoring calls for more accomplished analysis mechanisms than the typical text analysis tools provide today. The notion of word space is introduced, and it is argued that word spaces can be used to select the most salient markers for topicality, find associations those observations engender, and that they constitute an attractive foundation for building a representation well suited for the tracking and monitoring of mentions of the entity under consideration.

1 Buzz monitoring as a text analysis task

Buzz monitoring is the task of tracking text sources, with special attention given to user- and consumer-generated discussions, for mentions of some particularly interesting textual entity — a product, a service, a brand or similar — on the Internet. The task has gained widespread attention both as an interesting research excercise and as a useful and practical application [1]. It is widely understood that word of mouth phenomena play an important role for informing and affecting consumer decisions, and in building and destroying brand reputation. User-generated Internet content such as forums, blogs, and BBS's facilitate this process, and are taking over the authoritative status historically bestowed on traditional media, especially in markets where the authority and independence of traditional media is low for political or commercial reasons. Marketing strategists are increasingly becoming aware of the importance of "the buzz".

Similarly, intelligence and security analysts want to identify and keep track of certain user-initiated discussions and postings on forums, blogs, newsgroups, and other user generated web content. Albeit considerably more complex, the intelligence and security task is parallell to buzz monitoring, and pose similar demands on the analysis tools employed. Besides the fact that attitude analysis itself is complex for traditional knowledge-intensive methods, moving from traditional textual data to user-contributed content involves new and substantial challenges for text analysis methods.

New text as an object for analysis

Recent advances in publication and dissemination systems have given rise to new types of text — dynamic, reactive, multi-lingual, with numerous cooperating or even adversarial authors. Many of these new types of text remain true to established existing textual genres and conform to standard usage; others break new ground, moving towards new emergent textual genres enabled by the dramatically lowered publication threshold and distribution mechanisms.

Most notably these new forms of text, with a considerable amount of attention from traditional media, include forums, blogs, and wikipedias built as timely running commentary of public or private matters, individually and cooperatively. While heterogenous as a category, these new text forms share features in that they are subject to little or no editorial control compared to traditional media with higher publication thresholds. This makes the language used in new forms of text much more likely to pose complex technical challenges to traditional text analysis tools and mechanisms. New text bridges — in many cases — some of the character of written and spoken language, with rapid topical shifts, ad-hoc usage and coinage of terms, and a high degree of anchoring in the discourse context: to understand an utterance, one must understand the context it has been uttered in.

Human language does not lend itself easily to objectively trustworthy analysis of topical mention. A number of characterics of language in general and specifically its usage in new text combine to make text analysis non-trivial:

- Human usage of language is based on individual and social factors, not necessarily accessible to precomputed notions of meaning. Terms that have or gain some specific meaning in some social context do not necessarily carry that same meaning in other contexts, especially if the cultural remove between social contexts is great. The associative power of terms is an important component of meaning, but nigh impossible to formalise and model reliably in static knowledge sources.
- Polysemy, vagueness, and indefiniteness are all important and necessary characteristics of human communication: words are only cues for topical comprehension and variable value for that purpose. In fact, there is no exact matching from words to concept. Words are besides vague both polysemous and incomplete: every word means several things and every thing can be expressed with any one of several words. Words suffer from the combined drawback of being ambiguous and non-exclusive as indicators of content.
- The temporal character of information streams necessitate a dynamic model of topical focus, in contrast with most well-established models of topical analysis such as are applied in search engines and other information access contexts. As in the case of social contexts, one discourse situation to another the usage and prototypical referents of expressions shift and change with little or no confusion for human users; as time passes, words' meanings evolve and change with little or no confusion, without any attention from their users. Word meaning can be established or redefined during the course of one single conversation, with scope varying from local to universal.

Thus, any model intended to work with multiple and new text sources must have readiness to realign its semantic models with little forewarning. Traditional lexical resources may be useful as a base, but for the effective and reliable analysis of dynamic streams of texts a learning component is essential.

This paper argues that the notion of *word space*, which is a general framework for modelling semantic relations between terms in usage, presents an attractive alternative to standard text analysis mechanisms, and that it is especially suited for the tracking and monitoring of mentions of some topically interesting textual entity — say a trade mark, some identified entity, or a location — from a continuously and dynamically changing text collection, a news feed, the output from an Internet crawler or similar text source. In the following sections, we first review frequency-based mechanisms for topical text analysis and discuss why they are insufficient for buzz monitoring purposes. We then present the notion of word space, and discuss how it can be used for selecting the most salient markers for topicality, find associations those observations engender, and for building a representation well suited for the tracking and monitoring of mentions of the entity under consideration. We also discuss a potential problem with the proposed approach, and argue that the *Random Indexing* word space methodology can be used to overcome this problem that is inherent in other word space approaches.

2 Counting words as indicators of document topic

Information analysis systems view documents as carriers of topical information, and hold words and terms as reasonable indicators of topic. The idea is that we can get a useful indication of what is being talked about in a document if we look at which words are used in it. If, for example, the terms "sarin" and "attack" show up in a particular data source, we have a good clue what the topic is. By the same token, if the words "priceworthy" and "recommend" turn up in data discussing a particular product, we can make a qualified guess that the product is discussed in quite favorable terms.

The basic assumption of automatic indexing mechanisms is that the presence or absence of a word — or more generally, a term, which can be any word or combination of words — in a document is indicative of topic. This is obviously a simplistic view of topic and text;[3] many words are mentioned in passing; many terms used metaphorically; quoting may confuse the simplest mechanisms. Choice of which word occurrences to make note of and which to disregard is the central task for the accomplished text analysis computation mechanism.

Most systems in use today take the frequency of observed occurrence as a proxy for topicality. The first base assumption, originally first formulated by Hans Peter Luhn [2], is that infrequent words are uninteresting, being mostly non-topical or near-random mentions, and that very frequent words are structural and functional rather than topical (cf. Fig 1). The second base assumption is that the use of words elsewhere in the collection under consideration is a fair

[3] Written text, or other forms of discourse. We work mainly on written sources for the purposes of this discussion.

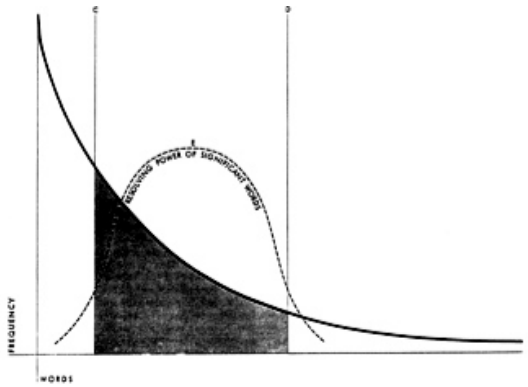

Fig. 1. Word-frequency diagram: X-axis represents individual words arranged in order of frequency. (From [2]).

source of knowledge as to their topical centrality and discriminatory power in the text under consideration [3]. Exactly how the suitably frequent terms are winnowed out using these two measures of salience and specificity is an algorithmic question with a number of parametrized formulæ to choose from for implementation.

As an example of what frequency counting can give us, we have computed term frequencies for a number of blog texts recently collected from various blog sources ($\approx$ 100) mentioning the Sony trade mark. Table 1 contrasts content terms that are frequent (left column) with content terms that are unexpectedly frequent (right column) as compared with a background text collection of similar texts on other topics ($\approx$ 15 000). The words shown in the right column are chosen by their observed term frequency showing a marked difference to expected term frequency as computed by χ^2, a standard non-parametric distribution test.[4] It is clear that a refined model provides a better basis for analysis: if we had not known this before, a simple χ^2 analysis would afford us license to assume that the Sony texts have something to do with technology, most likely cellular technology. But no terms indicating attitude or affect are even close to the top hundred or so terms.

More advanced methods for selecting and promoting salient terms can be utilized in order to avoid considering every occurring word as a candidate for inclusion in the analysis. Examples include selecting terms that are emphasized or repeated, selecting multi-word terms, finding variants or inflected forms of terms, using syntactic cues to identify terms that are promoted to topical centrality, or finding terms that occur in more topical regions of the text.

The term selection methods discussed in this section are all tried and tested methods for document analysis, and constitute the backbone of most commer-

[4] $\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$, where O_i is an observed frequency, E_i is an expected frequency asserted by a null hypothesis, and n is the number of possible outcomes of the event.

Frequency	χ^2
still	philips
totally	picture
trade	sr100
pics	w850
think	sell
want	cellular
home	nokia
guys	samsung
thing	ericson
vaio	trade
phone	vaio
people	forum
forum	galaxy
station	ericsson
sony	sony

Table 1. Terms selected by term frequency (left column) and χ^2 (right column) from blog texts mentioning Sony.

cial information access systems today. Despite their apparent simplicity, few more sophisticated techniques manage to outperform word counting methods in standardized information access test settings. However, in buzz monitoring and intelligence and security analysis settings, we do not only want to have a vague indication of the topic being discussed, but also *how* the topic is being discussed — the attitude or affect in the text. Furthermore, it is normally the case that we do not know *exactly* which documents to look at; usually we know which data sources are of interest — like a particular URL or news feed — but we do not know which of the documents from that source (like individual blog postings, forum discussions, or news texts) are relevant. It is also not uncommon that we only have a vague idea about which sources to analyze. In such explorative scenarios, the word counting methods outlined above are obviously less useful.

3 Relating words in word space

In scenarios where we know which terms we are interested in — our *targets* — but where data is variant and noisy, we can utilize methods that relate terms to each other. Thus, rather than looking at individual term distributions as we do when we collect frequency information, we can relate such distributions to each other over the entire text collection. By doing so, we can model which other words are related to our targets in the data we are currently looking at, without having to identify exactly which documents are relevant for the targets.[5]

[5] Note that identifying relevant documents is a non-trivial task — some documents in which a target occurs will be irrelevant, and some documents will be relevant even if they do not mention the target at all.

A standard approach for modeling relatedness between words in information access research is to compute the contextual agreement between words over large text data. The idea is that if we find words — e.g. "VX" and "Novichok" — that tend to occur in the same contetxs — say, in the vicinity of "gas" — then we can assume that they are related to each other (this is based on the so-called *distributional hypothesis* of meaning [4]). Models of distributional similarity represent words as vectors $\mathbf{v}$ of occurrence frequencies:

$$\mathbf{v_i} = [f_j, \cdots, f_n]$$

where f is the frequency of (co-)occurrence of word i in (or with) context j. Such vectors are refered to as *context vectors*, since they represent the contexts in which a word has occurred. The contexts can be either other words or text regions. An example of the former kid of word space model is HAL (Hyperspace Analogue to Language [5]), and an example of the latter is LSA (Latent Semantic Analysis [6]). Different weighting schemes, thresholdings, and dimensionality reduction techniques like principal component analysis or singular value decomposition are then normally applied to the context vectors before similarity is computed by measuring the distance or angles between the resulting vectors. Since these models favor the use of linear algebra as implementational framework, they are normally refered to as *semantic spaces* or *word space models* [7].

This kind of models are extremely useful for automatic semantic analysis, and have been used for an impressive number of cognitive modeling scenarios and information access tasks, ranging from modeling vocabulary acquisition [6], word categorization [8] and lexical priming [9], to improving information retrieval [10], text categorization [11], knowledge assessment [12] and bilingual lexicon construction [13].

However, word spaces are often poorly understood and sometimes even misused. The potential problem is that the semantic content of the word spaces is defined by the kind of contexts that are used to construct them. As [4] shows, using different kinds of contexts leads to word spaces with different semantic content. In short, using words as contexts (as in the HAL model) leads to word spaces where words are related because they have a semantic connection (e.g. "attack" and "assault"), whereas using text regions as context (as in the LSA model) leads to word spaces where words are related because they have an associative relationship (e.g. "attack" and "chemical"). We will refer to the former type of model as a *semantic* word space, and to the latter as an *associative* word space.

When applying word spaces to buzz monitoring and intelligence and security analysis, this difference needs to be properly understood, since otherwise the analysis might be flawed or even misleading. For example, if our target is "playstation" and we are interested in other terms used to refer to this product, we will get different results depending on whether we are using a semantic or associative word space; the former space will give us words that are used *in the same way* as "playstation," such as "psp" — which is probably what we want — while the latter will give us words that are used *together with* "playstation," such as "sony" — which is part of what we already know when we set out to perform our analysis in the first place.

3.1 Terminology mining using word spaces

As hinted at in the previous section, we can use a semantic word space model to identify words that are used in similar ways in the data at hand — in effect, constructing a data specific lexicon. This is very helpful in particular when working with user-generated data because Internet slang and spelling variations are in abundance. Table 2 demonstrates a number of words that a semantic word space model found to be related to the target word "recommend" in a large collection of blog data. The similarity column indicates the degree of relatedness (computed as the cosine of the angles between context vectors) and the type column specifies the type of relation between the words. As can be seen, the first three related words are spelling variations of the target word, and the five last words are spelling variations to a domain specific synonym (in this blog data, "love" is often used synonymously with "recommend").

Table 2. Words related to "recommend" in a semantic word space.

Related word	Similarity	Type
"recomend"	0.972	spelling variation
"reccomend"	0.968	spelling variation
"reccommend"	0.941	spelling variation
"looove"	0.870	spelling variation for "love"
"loooove"	0.863	spelling variation for "love"
"lurve"	0.850	spelling variation for "love"
"love"	0.846	the correct spelling of "love"
"loooooove"	0.836	spelling variation for "love"

Many of these relations would not have been previously known by a human analyst (e.g. the connection between "recommend" and "lurve"), and would only be detected by consulting a semantic word space model built from actual blog data; it is obvious that the majority of spelling variations are practically impossible to foresee. Furthermore, Internet slang and domain specific terminology may pose severe problems for human analysts. This is particularly true for intelligence and security analysis, where subjects are likely to consciously use opaque and even secretive terminology.

3.2 Opinion mining

Associative word spaces can also be very useful tools for buzz monitoring and intelligence and security analysis. Recall from Section 3 that in these types of word spaces words are close to each other if they have been used together in the data. This makes associative word spaces suited to use for opinion mining, where the task is to find out *how* subjects talk about a specific target. As we saw in

Section 2, listing the most frequent terms in documents mentioning the target does not say very much about the attitude expressed in the documents towards the target. However, by computing the similarity between a target and an attitudinally loaded term in an associative word space, we can get an indication of how the target is being talk about in the data.

The idea is to construct a number of pre-determined poles of interest in the associative word space, and to relate the target(s) to these poles. Such poles can be anything of interest for the analyst (threat, risk, stability, quality, reliability, sexiness, etc.). Figure 2 demonstrates the idea; "MyProduct" is situated closer to "Good" than to "Bad," which indicates that "MyProduct" has been talked about more often in terms of "good" than in terms of "bad."

Fig. 2. Opinion mining using poles in associative word space.

We introduced this technique in [14], in which a number of short news headlines were annotated with emotional valence using an associative word space model and a number of seed terms expressing bad versus good valence. The word space was built from a corpus of US newsprint available for experimentation for participants in the Cross Language Evaluation Forum (CLEF).[6] Eight negative and eight positive seed words were then used to construct one negative and one positive pole in the word space by simply taking the centroid of the seed word vectors:

$$\mathbf{v}_S = \sum \mathbf{v}_{w \in S}$$

where S is one of the seed sets, and w is a word in this set. Each news headline was similarly expressed as a centroid of the composing words after lemmatization and stop word filtering. The centroid headline vector was then compared to each of the pole vectors by computing the cosine of the angles between the vectors, and the pole whose vector had the highest cosine score was chosen to annotate the headline. The results were promising, and in particular generated very good recall.

[6] http://www.clef-campaign.org/

Applying this idea to our lage collection of blog data, Table 3 illustrates how a few different car brands relate to different poles in an associative word space. In the first example, "Volvo" and "Saab" are related to the pole "resale," and the similarity score shows that "Volvo" is much more related to "resale" than "Saab" is. The reason for this difference is that "Volvo" is discussed more in terms of "resale" than "Saab" in the blog data we analyzed, indicating that — in this data — Volvo's might be perceived as having a higher resale value than Saab's. Similarly in the second example, "Nissan" is much more related to the "good" pole than to the "bad" one, indicating that "Nissan" has a fairly positive valence in the particular blog data we analyzed.

Table 3. Examples of targtes related to poles.

Word and pole	Similarity
"volvo" → "resale"	0.348
"saab" → "resale"	-0.089
"nissan" → "good"	0.500
"nissan" → "bad"	-0.049

This general approach can be applied to any kind of opinion analysis task where it is possible to define a number of poles in advance. In particular, we expect this technique to be useful for intelligence and security analysts, who typically have a very pronounced idea of which poles of interest are relevant for analysis.

4 Temporal change and word space

A particularly interesting aspect of buzz monitoring is to identify and track changes in the buzz. If a product that previously enjoyed a solid reputation in consumer-generated media is being discussed in very negative terms, this is likely to be something the marketing department would like be alerted about promptly. For intelligence and security analysis, temporal changes are typically even more important, and constitute the main locus of interest.

Unfortunately, temporal changes can be hard to interpret when merely analyzing frequency-based lists of words; this requires considerable manual analysis, and is prone to error. Word spaces, on the other hand, provide a simple mechanism to detect changes in the buzz: if a target drifts in relation to the given poles in an associative word space, this is likely to indicate the presence of some noteworthy process. Figure 3 illustrates the idea. In this example, "MyProduct" has drifted from "Good" to "Bad" in our associative word space, which indicates a change in opinion among our costumers.

A potential problem with handling temporal change in word spaces is that many word space techniques are not designed to handle dynamic data. That is,

Fig. 3. Detecting change in associative word space.

when the data grows larger, so does the dimensionality of the word space, and it will sooner or later become computationally intractable. Many word space implementations therefore use statistical dimensionality reduction techniques to handle the problem of high dimensionality. A very common choice is singular value decomposition, which is the hallmark of Latent Semantic Analysis. Unfortunately, the computational cost of singular value decomposition is prohibitive, and once it has been performed it is non-trivial to add new data to the word space. As such, methods based on LSA is a suboptimal choice for the dynamic task of buzz monitoring.

A more informed solution to the problem of increasing dimensionality is to use the Random Indexing methodology [15], in which the word space has a predetermined dimensionality that never increases, and in which new data can be added incrementally. This allows for handling of dynamic data streams, as well as analysis of temporal changes in the word space. Random Indexing works by using fixed-width vectors in which the (co-)occurrence counts of a specific context are recorded by several randomly chosen vector elements, thus constituting a distributed representation. Every time we encounter a word, we increment the random configuration of vector elements that represent the context(s) it occurs with or in, and if we encounter a new kind of context we simply assign it a new random configuration of vector elements. Thus, Random Indexing is an inherently incremental methodology that is also very efficient and extremely scalable. For details on Random Indexing, see [16]

5 A knowledge representation suited to specific tasks

The implementation of the word space model we work with is based on Random Indexing, and is specifically intended to address dynamic data and scenarios. Our model tasks are twofold — one is intended to capture an exploratory mode whereas the other is intended to track the closeness of some target to some pole of interest. We also want our model to be computationally tractable in face of a constant influx of relevant data — given the insight that prefiltering of data risks lowering the recall effectiveness of our application.

exploratory task: What is X related to these days?
monitoring task: How closely are X and Y associated?

A typical feature of non-formal discourse is its dynamic nature — usage is coined on the fly, terms appear and disappear, mentions change. This both reflects the nature of language itself, but also, more crucially for our present application purposes, the changing topical associations under treatment: when the linguistic target obtains new associative terms or loses previously central associative neighbours, this is noteworthy. For this purpose our implementation of the word space model includes a model of change that can indicate recent changes in the semantic neighborhood of a term, and in the association between a target and pole.

This paper has argued that word spaces provide an attractive general framework to a problem which traditional information access models will have trouble handling. However, using word spaces require an understanding of how context models can best be parametrized to yield the most meaningful relations for the task at hand — e.g., as indicated above, noting the difference between semantic and associative word spaces; they also are likely to prove intractable in practice for large amounts of incoming data unless implemented using incremental learning models such as Random Indexing. Our implementation[7] is built to handle both the fine-grained distinctions of various contextual models and the demands posed by large scale incremental learning.

References

1. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Foundation and Trends in Information Retrieval **2**(1-2) (2008) 1–135
2. Luhn, H.: The automatic creation of literature abstracts. IBM Journal of Research and Development **2**(2) (1958) 159–165
3. Jones, K.S.: A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation **28** (1972) 11–20
4. Sahlgren, M.: The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces. PhD Dissertation, Department of Linguistics, Stockholm University (2006)
5. Lund, K., Burgess, C., Atchley, R.: Semantic and associative priming in high-dimensional semantic space. In: Proceedings of the 17th Annual Conference of the Cognitive Science Society, CogSci'95, Erlbaum (1995) 660–665
6. Landauer, T., Dumais, S.: A solution to plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. Psychological Review **104**(2) (1997) 211–240
7. Schütze, H.: Word space. In: Proceedings of the 1993 Conference on Advances in Neural Information Processing Systems, NIPS'93, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc. (1993) 895–902
8. Jones, M., Mewhort, D.: Representing word meaning and order information in a composite holographic lexicon. Psychological Review **114**(1) (2007) 1–37

[7] `http://www.gavagai.se`

9. McDonald, S., Lowe, W.: Modelling functional priming and the associative boost. In: Proceedings of the 20th Annual Conference of the Cognitive Science Society, CogSci'98. (1998) 675–680
10. Deerwester, S., Dumais, S., Furnas, G., Landauer, T., Harshman, R.: Indexing by latent semantic analysis. Journal of the Society for Information Science **41**(6) (1990) 391–407
11. Sahlgren, M., Cöster, R.: Using bag-of-concepts to improve the performance of support vector machines in text categorization. In: Proceedings of the 20th International Conference on Computational Linguistics, COLING'04. (2004) 487–493
12. Wolfe, M., Schreiner, M., Rehder, B., Laham, D., Foltz, P., Kintsch, W., Landauer, T.: Learning from text: Matching readers and text by latent semantic analysis. Discourse Processes **25** (1998) 309–336
13. Sahlgren, M., Karlgren, J.: Automatic bilingual lexicon acquisition using random indexing of parallel corpora. Journal of Natural Language Engineering **11**(3) (2005) 327–341
14. Sahlgren, M., Karlgren, J., Eriksson, G.: Valence annotation based on seeds in word space. In: Proceedings of Fourth International Workshop on Semantic Evaluations (SemEval'07). (2007)
15. Kanerva, P., Kristofersson, J., Holst, A.: Random indexing of text samples for latent semantic analysis. In: Proceedings of the 22nd Annual Conference of the Cognitive Science Society, CogSci'00, Erlbaum (2000) 1036
16. Sahlgren, M.: An introduction to random indexing. In Witschel, H., ed.: Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE'05, Copenhagen, Denmark, August 16, 2005. Volume 87 of TermNet News: Newsletter of International Cooperation in Terminology. (2005)