Quantitative Portraits of Lexical Elements

Kyo Kageura
Human and Social Information Research Division
National Institute of Informatics,
2-1-2 Hitotsubashi, Chiyoda-ku,
Tokyo, 101-8430, Japan
kyo@nii.ac.jp

Abstract

This paper clarifies the basic concepts and theoretical perspectives by and from which quantitative "weighting" of lexical elements are defined, and then draws, quantitative portraits of a few lexical elements in order to exemplify the relevance of the concepts and perspectives examined.

1 Introduction

Since Luhn's pioneering work (Luhn, 1958) in automatic term weighting, many methods have been proposed in the fields of IR (e.g. Spark-Jones, 1973; Harter, 1975) and NLP (e.g. Church et al., 1990). Some "standard" methods of term weighting such as $tfidf$ have been established (Aizawa, 2003; 徳永, 1999) and the application range has widened; term weighting has become a mature technology.

Despite this, what has been technically proposed has not been examined from a theoretical point of view, i.e. what kind of weighting scheme reflects what kind of lexical nature within what kind of framework of interpretations in language. We will clarify this and then illustrate the relevance of this clarification by drawing quantitative portraits of some lexical items using the quantitative measures.

2 Texts and lexica

Automatic term weighting starts from texts/documents. To what spheres the weights are attributed can differ. Figure 1 shows the linguistic spheres of lexica and texts (Kageura, 2002); there are both concrete data spheres and abstract spheres on both the lexical and textual sides.

Within this scheme, three types of relations between lexica and texts can be identified: concrete terms attributed to concrete texts, concrete terms corresponding to discourse, and abstract lexica corresponding to abstract discourse. We will show below that three major types of automatic term weighting methods correspond to these three types of relations between lexica and texts.

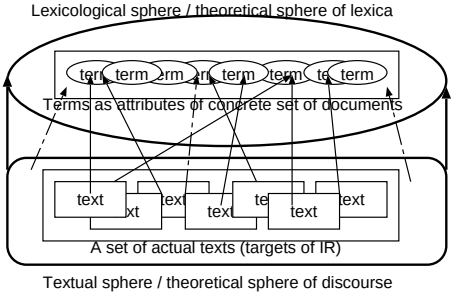

Figure 1: Textual sphere and lexicological sphere.

3 Methods of term weighting

3.1 Tfidf

$Tfidf$ is defined as:

$$tfidf = f_{t_i} \log \frac{N}{N_i} \quad (1)$$

where f_{t_i} is the total frequency of a term t_i, N is the total number of the documents, and N_i is the total number of documents in which the term t_i occurs.

Aizawa (2003) has shown that this can be derived from an information theoretic measure. Let $\mathcal{D}$ and $\mathcal{T}$ be random variables defined over events in a set of documents $D = \{d_1, d_2, ..., d_i, ..., d_N\}$ and a set of different terms $T = \{t_1, t_2, ..., t_j, ..., t_M\}$ in D. Let f_{ij} denote the frequency of t_i in d_j, f_{w_i} the total frequency of t_i, f_{d_j} the total number of running terms in d_j, and F the total number of term tokens in D. The "weight" of a term t_i can be given by:

$$\begin{aligned} \mathcal{F}(t_i; \mathcal{D}) &= P(t_i)\, \mathcal{K}(P(\mathcal{D}|t_i)||P(\mathcal{D})) \\ &= P(t_i) \sum_{d_j \in D} P(d_j|t_i) \log \frac{P(d_j|t_i)}{P(d_j)} \end{aligned}$$

Giving probabilities by relative frequencies, and assuming that all the documents have equal size and the frequency of t_i in the documents that contain t_i is equal, this measure becomes $tfidf$; $tfidf$ has an information theoretic meaning *within the given set of documents* (Figure 2).

3.2 Term representativeness

Hisamitsu, et al. (2000a) proposed a measure of "term representativeness", in order to overcome the

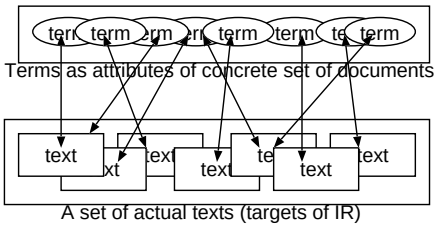

Figure 2: The position of $tfidf$.

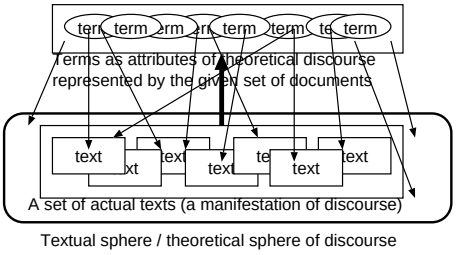

Figure 3: The position of term representativeness.

excessive sensitivity of weighting measures to token frequencies. They hypothesised that, for a term t, if the term is representative, D_t (the set of all documents containing t) have some specific characteristic. They define a measure which calculates the distance between a distributional characteristic of words around t and the same distributional characteristic in the whole document set.

In order to remove the factor of data size dependency, Hisamitsu et al. (2000a) defines the "baseline function," which indicates the distance between the distribution of words in the original document set and the distribution of words in randomly selected document subsets for each size. The distance between the distribution of words in the original document set and the distribution of words in the documents which accompany the focal term t is normalised by the "baseline function."

Formally,

$$Rep(t) = \frac{Dist(P_t, P)}{Dist(P_{R_t}, P)} \qquad (2)$$

where D denotes the set of all documents; P the distribution of words in D; t a focal term; D_t the set of all documents containing t; P_t distribution of words in D_t; P_{R_t} distribution of words in randomly selected documents whose size equals D_t; $Dist(P_i, P_j)$ the distance between two distributions of words P_i and P_j. Log-likelihood ratio was used to measure the distance.

This measure observes the centripetal force of a term vis-à-vis discourse. i.e. *it captures the characteristic of terms in the general discourse as represented by the given set of documents* (Figure 3).

3.3 Lexical productivity

Nakagawa (2000) incorporates a factor of lexical productivity of constituent elements of compound

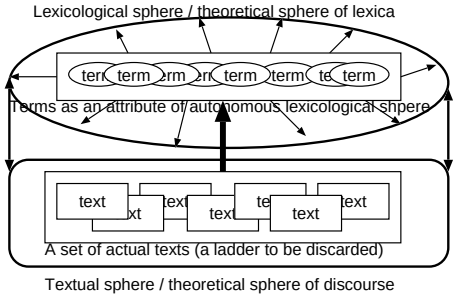

Figure 4: The position of lexical productivity.

units for complex term extraction. The method observes in how many different compounds an element t_i is used in a given document set (let us denote this as $d(i, N)$ where N indicates the size of the overall document set as counted by the number of word tokens), and used that in the weighting of compounds containing t_i, by taking weighted average. By explicitly limiting the syntagmatic range of observation of cooccurrence *to the unit of compounds*, he focused on the lexical productivity as manifasted in texts.

This measure depends on the token occurrence, but we can also think of the *theoretical* lexical productivity in the lexicological sphere: how many compounds t_i can *potentially* make" (let us denote this by $d(i)$). For that, it is necessary to remove the factor of token occurrence. This can be done by:

$$d(i) = d(i, \lambda N) \quad (\lambda \to \infty).$$

This has so far been unexplored. Potential lexical productivity of an element can be estimated from textual data: Letting p_{t_i} be the occurrence probability of t_i in texts, $f(i, N)$ be the token occurrence of t_i in texts, and C_i be the sample space $\{i_1, i_2, i_3, ..., i_{d(i)}\}$ of the distribution of compounds (and simplex word) that contains t_i with probability $p_{(c)i_k}$ given to each compound i_k, and assuming the combination of binomial distribution, we have:

$$\begin{aligned} \mathrm{E}[f(i,N)] &= p_{t_i} \cdot N \\ \mathrm{E}[d(i,N)] &= \sum_{m=1}^{p_{t_i} \cdot N} \sum_{k=1}^{d(i)} \binom{p_{t_i} \cdot N}{m} p_{i_k}^m (1 - p_{i_k})^{1-m}. \end{aligned}$$

What is given in the data is the empirical value for $d(i, N)$, with the empirical distributions of what actually occur in the document set among C_i. $d(i)$ can be estimated by LNRE methods (Baayen, 2001).

Being a measure representing the potential power of a lexical element t_i for constructing compounds, $d(i)$ indicates *the lexical productivity in the lexicological sphere* which correspond to theoretical sphere of discourse as represented by the given document set (Figure 4).

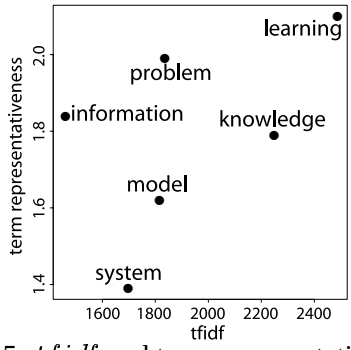

Figure 5: $tfidf$ and term representativeness.

4 Portraits of lexical elements

As the three different measures capture three different aspects of lexical elements, they are not competitive [1]. We here use these measures to illustrate characteristics of a few lexical elements.

We used NII morphologically tagged-corpus for observation (Okada et al., 2001), which consists of Japanese abstracts in the field of artificial intelligence. Table 1 shows the basic quantitative information.

No. of abstracts	word tokens (simplex/compound)	word types (simp./comp.)
1816	299846/230708	8764/23243

Table 1: The basic data for NII corpus.

We chose the six most frequently occurring nominal element for observation, i.e. システム (system), 知識 (knowledge), 学習 (learning), 問題 (problem), モデル (model), and 情報 (information). Intuitively, "system", and "model" are rather general with respect to the domain of artificial intelligence, "knowledge" and "learning" are domain specific, and "information" and "problem" are in between. Table 2 shows the basic quantitative information for these six lexical elements.

Figure 5 plots $tfidf$ and term representativeness for the six elements. Table 3 shows the estimated value of lexical productivity.

	p	$d(i)$
system	0.96	273402688337
knowledge	0.88	689
learning	0.39	2251563675
problem	0.70	1951
model	0.47	3676671255
information	0.84	667

Table 3: Lexical productivity for the six elements.

Figure 5 shows "learning" and "knowledge", intuitively the domain-dependent elements, take high $tfidf$ values, while "information" takes the lowest value. Term representativeness gives "learning" a high value but the values of "knowledge" is much lower, and about the same as "information". Interestingly, the lexical productivity of "knowledge" and "information" is also very close to each other.

It is possible to infer from these values of term representativeness and lexical productivity that both "information" and "knowledge" are, within the discourse of artificial intelligence, not with high centripetal value as both are rather "base" concepts of the domain. If we observe Table 2, "knowledge" is more often used as it is, while "information" tends to occur as compounds. From this we might be able to hypothesise that "knowledge" is in itself the "base" concept of artificial intelligence while "information" becomes the "base" concept in combination with other lexical items. This fits our intuition, as "information" in itself is more a "base" concept of information and computer science, which is a broader domain of which artificial intelligence is a subdomain. The low $tfidf$ value of "information" comes from the low token frequency coupled with relatively high DF, which shows that "information", as long as it is used, tends to scatter across documents. This is in accordance with the interpretation that "information" tends to occur in compounds. Still, however, it is difficult to interpret sensibly the fact that the $tfidf$ value of "information" is lower than those of "model" and "system". Perhaps it is more sensible to interpret $tfidf$ among elements which take the values of term representativeness higher than a certain threshold. Then we can say that "learning" and "knowledge" represent concepts more "central" to the domain of artificial intelligence than "information".

The element "learning", which takes the highest values both in $tfidf$ and in term representativeness, is conspicuous in its lexical productivity. Compared to "knowledge" whose $tfidf$ value is also high, and with the three elements "problem", "information" and "knowledge" whose term representativeness values are relatively high, the order of lexical productivity of "learning" is a million times higher (and similar to "model" or "system"). Table 2 shows that "learning" does not occur much as it is, nor does it occur much as the head of compounds. This indicates that "learning" represents an important concept of the given data and in the discourse of artificial intelligence, but only "indirectly" in combination with other elements in compounds where "learning" tend to contribute to as a modifier rather than a head.

The two "general" lexical elements, i.e. "model"

[1] It is thus simplistic to evaluate which measures work better in an application, unless the conceptual status of the applications is sufficiently clarified.

	TF	DF	Comp(A)	Comp(H)	Simp	$d(i,N)$(A)	$d(i,N)$(H)
system	2659	989	1922	1247	737	937	502
knowledge	2183	669	1399	443	784	424	137
learning	1776	462	1513	208	263	375	73
problem	1758	660	1197	558	561	334	152
model	1480	550	1144	687	343	447	263
information	1038	460	656	268	382	207	155

Note: Comp(A) indicates the number of compounds that contains the lexical element; Comp(H) indicates the number of compounds that contains the lexical element as the head; $d(i,N)$(A) indicates the number of different compounds (plus one simplex) that contains the lexical element; $d(i,N)$(H) indicates the number of different compounds (plus one simplex) that contains the lexical element as the head.

Table 2: The basic data for the six lexical elements.

and "system", take low term representativeness values[2]. This is in accordance with our intuition. The lexical productivity of these two elements are extremely high (practically infinite). This indicates that these two elements can be widely used in varieties of discoursal contexts, without in itself contributing much to consolidating the content of discourse. This fits nicely to our intuitive interpretation of the meanings of these two elements, i.e. they are orthogonal to to such domain-dependent elements as "knowledge" or "learning".

This leaves us with the final element "problem". The value of term representativeness is high, second only to "learning" and in between "learning" and "information"/"knowledge". The lexical productivity is much closer to "information" and "knowledge" than to the other three. As such, "problem" can be interpreted as a kind of "base" concept, though it retains stronger centripetal force than "information" and "knowledge". If we ignore $tfidf$ values of "model" and "system" and only compare "information", "problem", "learning" and "knowledge", it is also sensible to see that "problem" represent a concept more central to the domain than "information" but less central than "learning" and "knowledge".

5 Conclusions

We have shown that different term weighting measures have different spheres of interpretation; on the basis of that we have illustrated that the combination of these can illustrates complex aspects of lexical nature. Though it can be argued that the present study does not show ways for applications nor "empirical" evaluations within applications, we believe that "empirical" evaluations should be properly founded by the framework of interpretation in order for the results to be generalised in a scientific way; history of sciences have shown that often reliance on "empirical" evaluations correlates with the lack of theory or scientific wholesomeness.

[2]This is in accordance with the observation by Hisamitsu et al. (2000) which says that the measure of term representativeness is particularly useful to exclude general elements.

References

Akiko N. Aizawa. 2003. An information-theoretic perspective of tf-idf measures. *Information Processing and Management*, 39(1): 45–65.

Harald Baayen. 2001. *Word Frequency Distributions*. Dordrecht: Kluwer.

Kenneth W. Church and Patrick Hanks. 1990. Word association norms, mutual information and lexicography. *Computational Linguistics*, 16(1): 22–29.

S. P. Harter. 1975. A probabilistic approach to automatic keyword indexing. *Journal of the American Society for Information Science*, 26(4): 197–206.

Toru Hisamitsu, et. al. 2000. A method of measuring term representativeness. *COLING 2000*, 320–326.

Kyo Kageura. 2002. *The Dynamics of Terminology*. Amsterdam: John Benjamins.

Hans P. Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2): 159–165.

Hiroshi Nakagawa. 2000. Automatic term recognition based on statistics of compound nouns. *Terminology*, 6(2): 195–210.

Maho Okada, et. al. 2001. Defining principled but practically manageable lexical units in Japanese textual corpora. *NLPRS'01 Workshop on Language Resources in Asia*, 47–53.

Karen Sparck-Jones. 1973. Index term weighting. *Information Storage and Retrieval*, 9(11): 619–633.

徳永健伸. 1999. **情報検索と言語処理**. 東京: 東大出版会.