

# HANDLING THE IMPACT OF LOW FREQUENCY EVENTS ON CO-OCCURRENCE BASED MEASURES OF WORD SIMILARITY

## *A Case Study of Pointwise Mutual Information*

François Role and Mohamed Nadif

LIPADE, Université Paris Descartes, 45, rue des Saints Pères, 75006 Paris, France

Keywords: Text mining, Word similarity measures, Pointwise mutual information.

Abstract: Statistical measures of word similarity are widely used in many areas of information retrieval and text mining. Among popular word co-occurrence based measures is Pointwise Mutual Information (PMI). Although widely used, PMI has a well-known tendency to give excessive scores of relatedness to word pairs that involve low-frequency words. Many variants of it have therefore been proposed, which correct this bias empirically. In contrast to this empirical approach, we propose formulae and indicators that describe the behavior of these variants in a precise way so that researchers and practitioners can make a more informed decision as to which measure to use in different scenarios.

## 1 INTRODUCTION

The ability to automatically discover semantically associated words is at the heart of many information retrieval and extraction applications. For example, when querying a search engine, many terms can be used to express the same or very similar thing (e.g. "city" and "town"). As a consequence, some relevant documents may not be retrieved just because they do not contain exactly the same words as those in the query text. A common technique for overcoming this lack of flexibility is to expand an initial query using synonyms and related words. Being able to identify semantically similar words can also be useful in classification applications to reduce feature dimensionality by grouping similar terms into a small number of clusters. Many measures of word semantic similarity rely on word co-occurrence statistics computed from a large corpus of text: the assumption is that words that frequently appear together in text are conceptually related (Manning and Schütze, 1999; Lee, 1999; Thanopoulos et al., 2002; Evert, 2004; Pecina and Schlesinger, 2006; Bullinaria and Levy, 2007; Hoang et al., 2009). Among the most popular word co-occurrence based measures is Pointwise Mutual Information (PMI). For two words (or terms)  $a$  and  $b$ , PMI is defined as:

$$\log \frac{p(a,b)}{p(a)p(b)}$$

$p(a)$  (resp.  $p(b)$ ) is the probability that word  $a$  (resp.  $b$ ) occurs in a text window of a given size while  $p(a,b)$  denotes the probability that  $a$  and  $b$  co-occur in the same window. PMI is thus the log of the ratio of the observed co-occurrence frequency to the frequency expected under independence. It measures the extent to which the words occur more than by chance or are independent. The assumption is that if two words co-occur more than expected under independence there must be some kind of semantic relationship between them. Initially used in lexicography (Church and Hanks, 1990), PMI has since found many applications in various areas of information retrieval and text mining where there is a need for measuring word semantic relatedness. In (Vechtomova and Robertson, 2000), the authors show how to use PMI to combine a corpus information on word collocations with the probabilistic model of information retrieval. In the context of passage retrieval PMI is used to expand the queries using lexical affinities computed using statistics generated from a terabyte corpus (Terra and Clarke, 2005). Other examples include the use of PMI in projects devoted to extracting entities from search query logs, discovering indexing phrases, predicting user click behavior to simulate human visual search behavior, etc. Last but not least, automatic synonym discovery is another area where PMI has been extensively used (Turney, 2001; Terra and Clarke, 2003).

However, a well-known problem with PMI is its

tendency to give very high association scores to pairs involving low-frequency words, as the denominator is small in such cases, while one generally prefers a higher score for pairs of words whose relatedness is supported by more evidence.

The impact of the bias is apparent in tasks directly related to query transformation. For example, in (Croft et al., 2010), while talking about query expansion, the authors take a collection of TREC news as example, and show that, according to PMI, the most strongly associated words for "tropical" in this corpus are: "trmm", "itto", "ortuno", "kuroshio", "bio-function", etc. though the collection contains words such as "forest", "tree", "rain", "island" etc. They then conclude that these low-frequency words "are unlikely to be much use for many queries".

Nonetheless, the basic straightforwardness of PMI over other approaches is still appealing, and several empirical variants have therefore been proposed to overcome this limitation. Since the product of two marginal probabilities in the denominator favors pairs with low-frequency words, a common feature of these variants is to assign more weight to the joint probability  $p(a, b)$  either by raising it to some power  $k$  in the denominator ( $\log \frac{p(a, b)^k}{p(a)p(b)}$ ) or by using it to globally weight PMI as in the case of the so-called "Expected Mutual Information" ( $p(a, b) \log \frac{p(a, b)}{p(a)p(b)}$ ). However, as pointed out by (Croft et al., 2010), the correction introduced may result in too general words being top-ranked.

Whether it is preferable to discover specialized or general related terms depends on the context. In any case, the point is that failing to precisely quantify the impact of the bias and its possible corrections inevitably leads to empirical results that are very dependent on the data. The aim of this paper is therefore to propose precise indicators of sensitivity to frequency.

The plan of the paper is as follows. In section 2, we review PMI and some common variants in order to give insight into how each measure works, factors influencing them, and their differences. We also propose formulae for assessing the impact of the corrections brought by several widely used variants. Section 3 provides experimental validation of these formulae and investigates how to give some simple visual hints at the differences in behaviour to be expected when migrating from one measure to the other. We conclude by summarizing our contribution and indicating directions for future research.

## 2 A FORMAL STUDY OF SOME IMPORTANT VARIANTS OF PMI

Although widely used, PMI<sup>1</sup> has two main limitations : first, it may take positive or negative values and lacks fixed bounds, which complicates interpretation. Secondly, it has a well-known tendency to give higher scores to low-frequency events. While this may be seen as beneficial in some situations, one generally prefers a higher score for pairs of words whose relatedness is supported by more evidence.

In order to overcome these limitations, several variants of PMI have therefore been proposed over the years. In contrast to more general relatedness measures for which numerous comparative studies are available (Pecina and Schlesinger, 2006; Hoang et al., 2009; Thanopoulos et al., 2002; Lee, 1999; Petrovic et al., 2010; Evert, 2004), no systematic and formal comparison specifically addressing these variants seems to have been conducted so far.

Among the most widely used variants are those of the so-called PMI<sup>k</sup> family (Daille, 1994). These variants consist in introducing one or more factors of  $p(a, b)$  inside the logarithm to empirically correct the bias of PMI towards low frequency events. The PMI<sup>2</sup> and PMI<sup>3</sup> measures commonly employed are defined as follows:

$$\text{PMI}^2(a, b) = \log \frac{p(a, b)^2}{p(a)p(b)}$$

and

$$\text{PMI}^3(a, b) = \log \frac{p(a, b)^3}{p(a)p(b)}.$$

Note that from the expression of PMI<sup>2</sup>( $a, b$ ), a simple derivation shows that it is in fact equal to  $2 \log p(a, b) - (\log p(a) + \log p(b))$ , and thus to

$$\text{PMI}(a, b) + \log p(a, b).$$

That is to say the correction is obtained by adding some value increasing with  $p(a, b)$ , namely  $\log p(a, b)$  to PMI, which, in fact, will boost the scores of frequent pairs. However, for comparison purposes it may be more convenient to express PMI<sup>2</sup>( $a, b$ ) as:

$$\text{PMI}^2(a, b) = \text{PMI}(a, b) - (-\log p(a, b)) \quad (1)$$

<sup>1</sup>PMI is not to be confused with the Mutual Information between two discrete random variables  $X$  and  $Y$ , denoted  $I(X; Y)$ , which is the expected value of PMI.

$$I(X; Y) = \sum_{a, b} p(a, b) \log \frac{p(a, b)}{p(x)p(y)} = \sum_{a, b} p(a, b) \text{PMI}(a, b).$$

The transformation may seem purely cosmetic but it will allow us to better analyze the PMI variants in information-theoretic terms. Indeed, frequent word pairs convey less information (in information-theoretic sense) than infrequent ones. Therefore, when writing

$$\text{PMI}(a,b) - (-\log p(a,b)),$$

the reasoning is that the more frequent a pair, the less information we subtract from the PMI. Note that, a generalization of (1) is easily obtained by writing:

$$\text{PMI}^k(a,b) = \text{PMI}(a,b) - (-(k-1)\log p(a,b)).$$

This expression of  $\text{PMI}^k$  measures allows us to compare them in a precise way to the NPMI (Normalized PMI) recently proposed in (Gerlof, 2006). The main motivation for this new variant is to give PMI a fixed upper bound of 1 in the case of perfect dependence, that is to say when two words only occur together. In this case

$$\text{PMI}(a,b) = -\log p(a) = -\log p(b) = -\log p(a,b).$$

One option to normalize PMI is then to divide it by  $-\log p(a,b)$ , which results in the following definition:

$$\text{NPMI}(a,b) = \frac{\text{PMI}(a,b)}{-\log p(a,b)} \quad (2)$$

Another common variant consists of removing the log that appears in the definition of PMI. We then obtain a new positive variant of PMI noted PPMI and defined as:

$$\text{PPMI}(a,b) = 2^{\text{PMI}(a,b) - (-\log p(a,b))} \quad (3)$$

We eventually end up with expressions of the variants which exhibit what they have in common.

We conclude this review by comparing  $\text{PMI}^2$ , PPMI and NPMI, based on how their sensitivity to low frequencies is affected by the correction factors we have just described. In table 1, we first recap the values taken by the variants in three cases: complete dependence (both words only occur together), independence, and when  $p(x,y) = 0$ .

Table 1: Variation intervals of PMI, NPMI,  $\text{PMI}^2$  and PPMI.

Variants	Comp.	Ind.	null
PMI	$-\log p(a,b)$	0	$-\infty$
NPMI	1	0	-1
$\text{PMI}^2$	0	$\log p(a,b)$	$-\infty$
PPMI	1	$p(a,b)$	0

Let us now assess to what extent the correction factors used by the three different variants influence the scores assigned to the pairs. In order to quantify this influence of frequencies, it is easy to prove the following proposition concerning  $\text{PMI}^2$ :

**Proposition 1.**

Let  $k \geq 0$ , if  $\text{PMI}(c,d) = \text{PMI}(a,b) + k$ , then

$$\begin{cases} \text{PMI}^2(c,d) = \text{PMI}^2(a,b) & \text{if } \log \frac{p(a,b)}{p(c,d)} = k \\ \text{PMI}^2(c,d) > \text{PMI}^2(a,b) & \text{if } \log \frac{p(a,b)}{p(c,d)} < k \\ \text{PMI}^2(c,d) < \text{PMI}^2(a,b) & \text{if } \log \frac{p(a,b)}{p(c,d)} > k \end{cases}$$

Using the second implication, we can also show that  $\text{PMI}^2$  increases with  $p(a,b)$ . Indeed, in the case when  $k = 0$ , the implication says that if

$$\log \frac{p(a,b)}{p(c,d)} < 0$$

then

$$\text{PMI}^2(c,d) \geq \text{PMI}^2(a,b).$$

In addition, since  $\text{PPMI}(a,b) = 2^{\text{PMI}^2}$  the same implications can be used for PPMI.

In the same manner, in the following proposition, we can prove and therefore quantify the influence of the frequencies for the NPMI measure.

**Proposition 2.**

Let  $\text{PMI}(z,t) = k\text{PMI}(x,y)$  we have:

- \*  $\text{NPMI}(z,t) = \text{NPMI}(x,y)$ 
  - if  $\log p(z,t) = k \log p(x,y)$ .
- \*  $\text{NPMI}(z,t) > \text{NPMI}(x,y)$ 
  - if  $(k > 0, \text{PMI}(z,t) < 0 \text{ and } \frac{\log p(z,t)}{\log p(x,y)} < k)$ ,
  - or  $(k > 0, \text{PMI}(z,t) > 0 \text{ and } \frac{\log p(z,t)}{\log p(x,y)} > k)$ ,
  - or  $(k < 0 \text{ and } \text{PMI}(z,t) > 0)$ .
- \*  $\text{NPMI}(z,t) < \text{NPMI}(x,y)$ 
  - if  $(k > 0, \text{PMI}(z,t) > 0 \text{ and } \frac{\log p(z,t)}{\log p(x,y)} < k)$ ,
  - or  $(k > 0, \text{PMI}(z,t) < 0 \text{ and } \frac{\log p(z,t)}{\log p(x,y)} > k)$ ,
  - or  $(k < 0 \text{ and } \text{PMI}(z,t) < 0)$ .

### 3 NUMERICAL EXPERIMENTS

The  $k$  thresholds derived in the previous section allow us to predict the consequences of switching from one measure to another. Take for example the following pairs (ba=basidiomycota, champ=champignon) and

(cha=chanson, si=single)<sup>2</sup> extracted from the French Wikipedia corpus described and used later.

The marginal and joint probabilities, and the computed PMI and PMI<sup>2</sup> measures of each pair are reported in table 2. As we have

$$PMI(ba, champ) - PMI(cha, si) = 3.05 = k$$

and

$$\log \frac{p(cha, si)}{p(ba, champ)} = 3.16 > 3.05,$$

we observe that

$$PMI^2(cha, si) > PMI^2(ba, champ),$$

despite

$$PMI(cha, si) < PMI(ba, champ).$$

Table 2: Marginal, joint probabilities, PMI and PMI<sup>2</sup> for two pairs.

(a, b)	basidiomycota champignon	chanson single
$p(a, b)$	$3.70 \times 10^{-5}$	$3.31 \times 10^{-4}$
$p(a)$	$3.95 \times 10^{-5}$	$5.68 \times 10^{-4}$
$p(b)$	$7.21 \times 10^{-5}$	$3.72 \times 10^{-4}$
$PMI(a, b)$	13.66	10.61
$PMI^2(a, b)$	-1.05	-0.94

Given that (chanson, single) is a much more frequent pair than (basidiomycota, champignon) PMI<sup>2</sup> was able to correct the bias of PMI so as to assign a higher score to the more frequent pair. Continuing with the previous example, this time using PMI and NPMI, one can show that in this case,  $k = 1.28$  and  $\log p(basidiomycota, champignon) \geq 1.28 \times \log p(single, chanson)$ . This leads to  $NPMI(basidiomycota, champignon) = 0.92 \geq NPMI(single, chanson) = 0.91$ . The more frequent pair is ranked after the less frequent one. In contrast to PMI<sup>2</sup>, NPMI was not able to correct the bias of PMI.

To stress how dramatic an impact this divergence may have on the ranking of related terms, table 3 shows the most strongly associated words (obtained using a large set of Wikipedia titles and categories we used as a corpus) for 'football' according to PMI and PMI<sup>3</sup>, which is a commonly used variant of PMI. As can be seen, we end up with two very different lists. PMI tends to rank domain-specific words at the top of the list whereas the top-ranked terms for PMI<sup>3</sup> are much more general.

<sup>2</sup>Translations: basidiomycota=basidiomycota; champignon=mushroom; chanson=song; single=single.

Table 3: Most strongly associated words for 'football' ranked according to the scores assigned by PMI (left) and PMI<sup>3</sup> (right). PMI considers the pair "football", "midfielder" to be the most strongly connected pair whereas PMI<sup>3</sup> ranks the pair "football", "league" first.

midfielder	5.581	league	-19.840
midfielders	5.575	clubs	-20.667
cornerbacks	5.545	england	-20.915
goalkeepers	5.543	players	-21.326
safeties	5.530	season	-21.677
goalkeeper	5.529	team	-21.922
linebackers	5.475	college	-22.043
striker	5.4750	club	-22.244
defenders	5.408	national	-22.891
defender	5.270	managers	-23.142
quarterbacks	5.262	cup	-23.498

This suggests a more global way of looking at the levels of correction brought about by the different variants for a particular corpus. First select a large corpus of word pairs. For each association measure, compute the score for every pair and rank the pairs by their scores. Then plot the rankings against each other so as to see whether the measures agree on which pairs of words contain the most strongly associated words. In order to have a statistically significant set of word pairs, we parsed the XML dump of the French Wikipedia dated September 14, 2009 (5 Gb of text). After filtering, we kept about one million articles that had at least a Wikipedia category associated with them<sup>3</sup>. We then parsed the articles, and for each word appearing in a category name, we recorded its co-occurrences with all the other words also appearing in category name. After removing French grammatical stop words as well as words occurring in less than 100 documents, we eventually ended up with a set of 70 922 word pairs. We then ranked these pairs according to the different measures and then plotted the so obtained ranks against each other.

The scatter plot in figure 1 shows that, compared to PMI, NPMI only slightly improves the bias towards low frequency pairs. The graph follows a rough diagonal. In contrast, figure 2 shows that moving from NPMI to PMI<sup>3</sup> has a dramatic effect as far as the assigned ranks are concerned. By looking at this kind of graph one can get a visual insight on the global effects of switching from one variant to another.

To get a more precise view, it is useful to focus on the most frequent pairs. The plots in figure 3 and figure 4 compare the ranks assigned by NPMI, PMI<sup>2</sup> and PMI<sup>3</sup> to the 500 most frequent pairs in the corpus.

<sup>3</sup>Categories are short phrases that represent the main topics of an article. Usually several categories are assigned to a given article.

From figure 3, it is clear that these pairs are ranked among the first by  $PMI^3$  while the ranks assigned by NPMI are more variable. This is a confirmation of figure 2 and shows that using  $PMI^3$  to find words related to another word would certainly result in quite more general terms than if using NPMI.

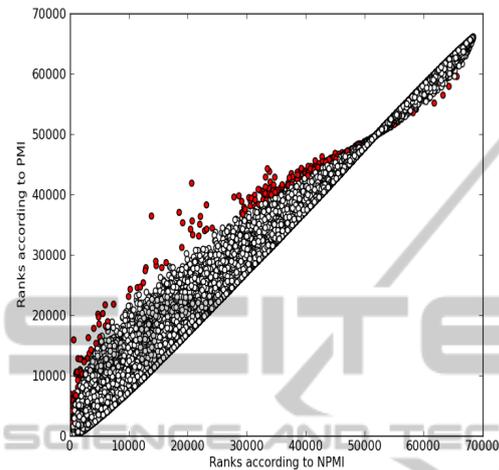


Figure 1: Scatter plot of the ranks assigned by PMI (vertical axis) and NPMI (horizontal axis). The plot reveals a near-linear relationship. The correction brought about by NPMI is not very significant for this corpus.

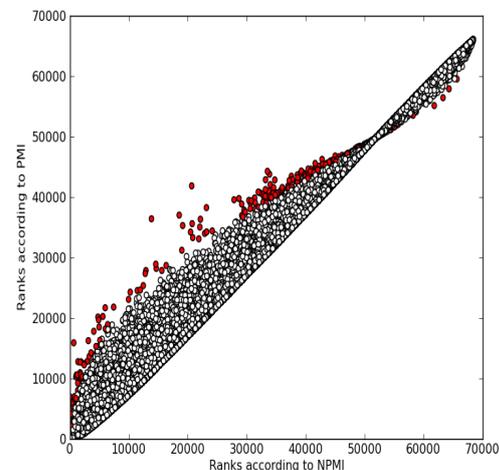


Figure 2: Scatter plot of the ranks assigned by NPMI (vertical axis) and  $PMI^3$  (horizontal axis).

#### 4 CONCLUSIONS

We have proposed formulae that help better understand and predict the behaviour of some important word association measures based on PMI, especially

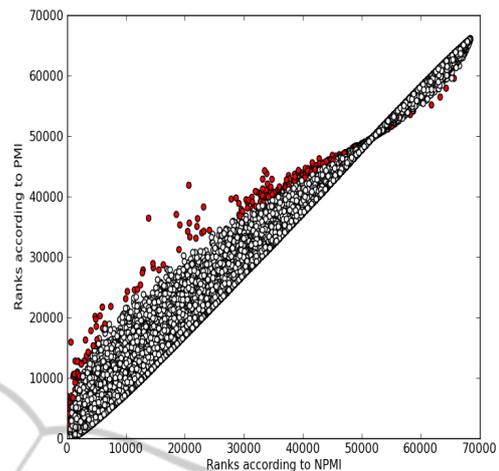


Figure 3: The 500 most frequent pairs as ranked by  $NPMI^3$  and  $PMI^3$ .

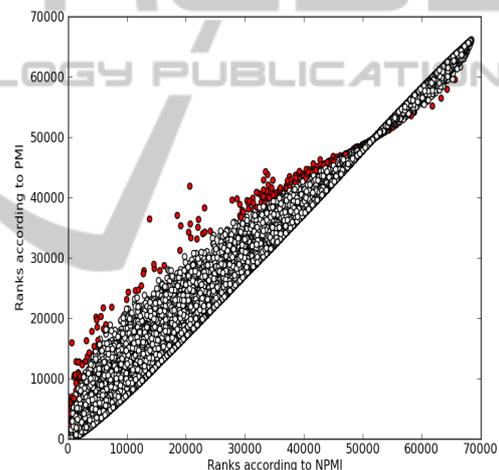


Figure 4: The 500 most frequent pairs as ranked by  $PMI^3$  and  $PMI^2$ . The most frequent pairs are clearly ranked first by both measures.

concerning the way these measures behaves w.r.t frequency. More specifically, to the best of our knowledge, it is the first time that the notorious bias of PMI towards low frequency and the ways to correct it have been examined in detail. We also feel that this study will help fill a gap in the literature by providing a comprehensive formalization and comparison of several important variants of PMI. In future, we plan to use rank correlation methods such as Spearman's Rho and Kendall's Tau to get another indicators of differences in sensitivity to frequency. Additional visualization methods will also be investigated in order to determine how to best help researchers and corpus practitioners in choosing the right PMI based association measure given the corpus and the task at hand.

## REFERENCES

- Bullinaria, J. and Levy, J. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, page 510.
- Church, K. W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.
- Croft, W. B., Metzler, D., and Strohman, T. (2010). *Search engines: information retrieval in practice*. Pearson - Addison Wesley.
- Daille, B. (1994). Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques. *PhD thesis, Universit Paris 7 (1994)*.
- Evert, S. (2004). The statistics of word cooccurrences: word pairs and collocations. *Unpublished doctoral dissertation, Institut für maschinelle Sprachverarbeitung, Universit ät Stuttgart*.
- Gerlof, B. (2006). Normalized (pointwise) mutual information in collocation extraction. In *Proceedings of the Biennial GSCL Conference*, pages 31–40, Gunter Narr Verlag, Chiarcos, Eckart de Castilho & Stede (eds).
- Hoang, H. H., Kim, S. N., and Kan, M.-Y. (2009). A re-examination of lexical association measures. In *MWE '09: Proceedings of the Workshop on Multiword Expressions*, pages 31–39, Morristown, NJ, USA. Association for Computational Linguistics.
- Lee, L. (1999). Measures of distributional similarity. In *37th Annual Meeting of the Association for Computational Linguistics*, pages 25–32.
- Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press.
- Pecina, P. and Schlesinger, P. (2006). Combining association measures for collocation extraction. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 651–658, Morristown, NJ, USA. Association for Computational Linguistics.
- Petrovic, S., Snajder, J., and Basic, B. D. (2010). Extending lexical association measures for collocation extraction. *Computer Speech & Language*, 24(2):383–394.
- Terra, E. and Clarke, C. L. A. (2003). Frequency estimates for statistical word similarity measures. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 165–172, Morristown, NJ, USA. Association for Computational Linguistics.
- Terra, E. and Clarke, C. L. A. (2005). Comparing query formulation and lexical affinity replacements in passage retrieval. In *SIGIR Workshop on Methodologies and Evaluation of Lexical Cohesion Techniques in Real-world Applications. Salvador, Brazil, August 2005*, pages 11–18. ACM Press.
- Thanopoulos, A., Fakotakis, N., and Kokkinakis, G. (2002). Comparative evaluation of collocation extraction metrics. In *Proceedings of the LREC 2002 Conference*, pages 609–613.
- Turney, P. D. (2001). Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. In *ECML-01*, pages 491–502, Freiburg, Germany.
- Vechtomova, O. and Robertson, S. (2000). Integration of collocation statistics into the probabilistic retrieval model. In *Proceedings of the 22nd British Computer Society - Information Retrieval Specialist Group Conference 2000*, pages 165–177.