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 Schutze, 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:

\[
\log \frac{p(a,b)}{p(a)p(b)}
\]

where \(p(a)\) (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: "tropical", "island", "rain", "function", etc. though the collection contains words such as "forest", "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 over other approaches is still appealing, and several

3 PROVIDES EXPERIMENTAL VALIDATION OF THESE FORMULAE (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: "TROPICAL", "ISLAND", "RAIN", "FUNCTION", ETC. THOUGH THE COLLECTION CONTAINS WORDS SUCH AS "FOREST", "RAIN", "ISLAND" ETC. THEY THEN CONCLUDE THAT THESE LOW-FREQUENCY WORDS "ARE UNLIKELY TO BE MUCH USE FOR MANY QUERIES".

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 family (Daille, 1994). These variants consist in introducing one or more factors of\[ p(a, b) \log \frac{p(a, b)}{p(a)p(b)} \] inside the logarithm to empirically correct the bias of PMI towards low frequency events. The PMI and PMI 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, 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 \( \text{PMI}^2(a, b) \) as:

\[
\text{PMI}^2(a, b) = \text{PMI}(a, b) - (- \log p(a, b)) \tag{1}
\]

1PMI 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.

\[
I(X;Y) = \sum_{a,b} p(a,b) \log \frac{p(a, b)}{p(a)p(b)} = \sum_{a,b} p(a,b) \text{PMI}(a,b).
\]
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:

\[ \text{PMI}^2(a, b) = \text{PMI}(a, b) - (-k) \log p(a, b). \]

This expression of PMI\(^2\) 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 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 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, PMI\(^2\) and PPMI.

<table>
<thead>
<tr>
<th>Variants</th>
<th>Comp.</th>
<th>Ind.</th>
<th>null</th>
</tr>
</thead>
<tbody>
<tr>
<td>PMI</td>
<td>(-\log p(a, b))</td>
<td>0</td>
<td>(-\infty)</td>
</tr>
<tr>
<td>NPMI</td>
<td>1</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>PMI(^2)</td>
<td>0</td>
<td>(\log p(a, b))</td>
<td>(-\infty)</td>
</tr>
<tr>
<td>PPMI</td>
<td>1</td>
<td>(p(a, b))</td>
<td>0</td>
</tr>
</tbody>
</table>

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 PMI\(^2\):

**Proposition 1.**
Let \(k \geq 0\), if \(\text{PMI}(c, d) = \text{PMI}(a, b) + k\), then

\[
\begin{align*}
\text{PMI}^2(c, d) &= \text{PMI}^2(a, b) & \text{if} \quad \log \frac{p(a, b)}{p(c, d)} &= k \\
\text{PMI}^2(c, d) &> \text{PMI}^2(a, b) & \text{if} \quad \log \frac{p(a, b)}{p(c, d)} &< k \\
\text{PMI}^2(c, d) &< \text{PMI}^2(a, b) & \text{if} \quad \log \frac{p(a, b)}{p(c, d)} &> k
\end{align*}
\]

Using the second implication, we can also show that 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 \((k > 0, \text{PMI}(z, t) < 0 \quad \text{and} \quad \frac{\log p(z, t)}{\log p(x, y)} < k)\),
  - or \((k > 0, \text{PMI}(z, t) > 0 \quad \text{and} \quad \frac{\log p(z, t)}{\log p(x, y)} > k)\),
  - or \((k < 0 \quad \text{and} \quad \text{PMI}(z, t) > 0)\).

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

### 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)² extracted from the French Wikipedia corpus described and used later. The marginal and joint probabilities, and the computed PMI and PMI² 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

PMI²(cha, si) > PMI²(ba, champ),

despite

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

Table 2: Marginal, joint probabilities, PMI and PMI² for two pairs.

<table>
<thead>
<tr>
<th>(a, b)</th>
<th>basidiomycota</th>
<th>champignon</th>
<th>chanson</th>
<th>single</th>
</tr>
</thead>
<tbody>
<tr>
<td>p(a, b)</td>
<td>3.70 × 10⁻⁵</td>
<td>3.31 × 10⁻⁴</td>
<td>3.05</td>
<td>1.28</td>
</tr>
<tr>
<td>p(a)</td>
<td>3.95 × 10⁻⁵</td>
<td>5.68 × 10⁻⁴</td>
<td></td>
<td></td>
</tr>
<tr>
<td>p(b)</td>
<td>7.21 × 10⁻⁵</td>
<td>3.72 × 10⁻⁴</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PMI(a, b)</td>
<td>13.66</td>
<td>10.61</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PMI²(a, b)</td>
<td>−1.05</td>
<td>−0.94</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Given that (chanson,single) is a much more frequent pair than (basidiomycota, champignon) PMI² 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 \( \log p(\text{basidiomycota, champignon}) \geq 1.28 \times \log p(\text{single, champignon}) \). This leads to NPMI(basidiomycota, champignon) = 0.92 ≥ NPMI(single, champignon) = 0.91. The more frequent pair is ranked after the less frequent one. In contrast to PMI², 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 PMI³, 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 PMI³ are much more general.

Table 3: Most strongly associated words for 'football' ranked according to the scores assigned by PMI (left) and PMI³ (right). PMI considers the pair "football", "midfielder" to be the most strongly connected pair whereas PMI³ ranks the pair "football", "league" first.

<table>
<thead>
<tr>
<th>Word</th>
<th>PMI</th>
<th>PMI³</th>
</tr>
</thead>
<tbody>
<tr>
<td>midfields</td>
<td>5.581</td>
<td>league</td>
</tr>
<tr>
<td>midfields</td>
<td>5.575</td>
<td>clubs</td>
</tr>
<tr>
<td>cornerbacks</td>
<td>5.545</td>
<td>England</td>
</tr>
<tr>
<td>goalkeepers</td>
<td>5.543</td>
<td>players</td>
</tr>
<tr>
<td>safety</td>
<td>5.530</td>
<td>season</td>
</tr>
<tr>
<td>goalkeeper</td>
<td>5.529</td>
<td>team</td>
</tr>
<tr>
<td>linebackers</td>
<td>5.475</td>
<td>college</td>
</tr>
<tr>
<td>striker</td>
<td>5.4750</td>
<td>club</td>
</tr>
<tr>
<td>defenders</td>
<td>5.408</td>
<td>national</td>
</tr>
<tr>
<td>defender</td>
<td>5.270</td>
<td>managers</td>
</tr>
<tr>
<td>quarterbacks</td>
<td>5.262</td>
<td>cup</td>
</tr>
</tbody>
</table>

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. 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 PMI³ 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, PMI² and PMI³ to the 500 most frequent pairs in the corpus.

²Translations: basidiomycota=basidiomycota; champignon=mushroom; chanson=song; single=single.
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).

Figure 3: The 500 most frequent pairs as ranked by NPMI 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.

4 CONCLUSIONS

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


Thanopoulos, A., Fakotakis, N., and Kokkinakis, G. (2002). Comparative evaluation of collocation extraction met-