Measuring Inter-Indexer Consistency Using a Thesaurus

Olena Medelyan
Department of Computer Science,
University of Waikato, Private Bag 3105
Hamilton, New Zealand
+64 7 838 4246
olena@cs.waikato.ac.nz

Ian H. Witten
Department of Computer Science,
University of Waikato, Private Bag 3105
Hamilton, New Zealand
+64 7 838 4246
ihw@cs.waikato.ac.nz

ABSTRACT
When professional indexers independently assign terms to a given document, the term sets generally differ between indexers. Studies of inter-indexer consistency measure the percentage of matching index terms, but none of them consider the semantic relationships that exist amongst these terms. We propose to represent multiple-indexers data in a vector space and use the cosine metric as a new consistency measure that can be extended by semantic relations between index terms. We believe that this new measure is more accurate and realistic than existing ones and therefore more suitable for evaluation of automatically extracted index terms.

Categories and Subject Descriptors
H.3.1 [**Content Analysis and Indexing**]: *Indexing methods.*

General Terms: Measurement, Documentation, Reliability.

Keywords: Inter-indexer consistency, controlled indexing.

1. INTRODUCTION
Indexing consistency has been defined as "the degree of agreement in the representation of the (essential) information content of a document by certain sets of indexing terms selected individually and independently by each of the indexers" [5]. Several different measures have been proposed, and many studies of inter-indexer consistency have been reported. They generally conclude that a high level of consistency is hard to achieve [1, 5] and that the indexers are more likely to agree on what concepts should be indexed than on the exact terms that best represent them [2, 3]. Surprisingly, existing consistency measures do not take into account the semantic relations that exist between terms in the indexing vocabulary, which intuitively would seem likely to improve accuracy.

2. MEASURING THE CONSISTENCY
There are two well known measures of inter-indexer consistency. Let A and B be the size of the two indexer's term sets and C be the number of terms in common between the two sets. Hooper's measure [1] is

$$\text{Hooper}(Indexer_1, Indexer_2) \quad H = \frac{C}{A + B - C}.$$

JCDL'06, June 11–15, 2006, Chapel Hill, North Carolina, USA.

Rolling's measure [4] is

$$\text{Rolling}(Indexer_1, Indexer_2) \quad R = \frac{2C}{A + B}.$$

Both range from 0 when the sets A and B are disjoint to 1 when they are identical. The two are related by $H = R/(2 - R)$, which shows that Hooper's measure is always smaller than Rolling's throughout the operating range [0,1]. Couched in the same terms, the cosine measure[1] can be expressed as:

$$\text{Cosine}(Indexer_1, Indexer_2) \quad \frac{C}{\sqrt{AB}}.$$

The sets can be represented as vectors, e.g. $\underline{A} = [A_1, A_2, \ldots, A_n]$, where n is the vocabulary of terms and the element A_i is 1 or 0 depending on whether term i is in the set A or not. Then

$$\text{Cosine}(Indexer_1, Indexer_2) \quad \frac{\sum_i A_i B_i}{\sqrt{\sum_i A_i^2}\sqrt{\sum_i B_i^2}} = \frac{\underline{A} \cdot \underline{B}}{|\underline{A}||\underline{B}|},$$

where $\underline{A} \cdot \underline{B}$ is the dot product of vectors. If the elements of the vectors are 0 or 1, their dot product is the number of elements they have in common. Given term sets from several different indexers, the single vector that represents their average can be used in the cosine measure to determine the similarity of an individual indexer to the group.

We obtained data from the UN Food and Agriculture Organization (FAO) in which 6 professional indexers independently assigned terms from the Agrovoc thesaurus (www.fao.org/agrovoc) to a set of 10 documents. Agrovoc contains 16,600 possible index terms and defines three semantic relations between them: bi-directional links between related terms (RT) and links between broader terms (BT) and narrower ones (NT), which are inverse.

The indexers assigned between 5 and 16 terms to each document, showing significant differences among each other. A surprisingly large number of assigned terms were idiosyncratic to a single indexer. Over the half of the assigned terms (150 of 280 different terms) were assigned to documents by a solitary indexer, and only 10 terms (3.6%) were agreed by all indexers. As a result, the average consistency among them is very low: 38% according to the Rolling's measure, and slightly better with the Cosine measure, 49.5%. However, the analysis of semantic relations among the assigned terms confirmed that although the terms do not match exactly, over the half of them are semantically related to other indexers' choices.

[1] The cosine measure uses the geometric mean of A and B in place of Rolling's arithmetic mean. Thus, the measures result in similar values, unless the sets unless have radically different sizes.

3. NEW MEASURE

It seems obvious that when comparing index sets A and B, one should take into account not just the terms they have in common, but terms in A that are related to some term in B and vice versa. Thus in our new consistency measure as well as equality we consider the two relations RT and BT/NT. We use numeric weights $\gamma, \alpha, \beta \in [0,1]$ to reflect the relative "importance" of these three effects. To ensure that the importance is relative, we make the weights sum to 1, and write $\gamma = 1 - \alpha - \beta$. It makes sense to demand that an increase in one weight necessarily involves decreasing the others; otherwise the measure of similarity could be raised artificially by simply increasing all weight values.

To take account of thesaurus relations, the similarity between two indexers A and B is estimated by computing the cosine measure between A's vector of terms $\underline{A}$ and a version $\underline{B}'$ of B's term vector that has been adjusted to reflect terms that are related to B's choices. First express the relations RT and BT/NT by $n \times n$ matrices $\underline{\underline{\mathrm{R}}}$ and $\underline{\underline{\mathrm{N}}}$ whose element at position i, j is 1 if term i is related to term j and 0 otherwise. Both these matrices are symmetric, the former because RT is a symmetric relation and the latter because it subsumes both the NT and BT relation, which are inverses. Then, using weights γ for identity and α and β for RT and BT/NT respectively, the adjusted version of B's term vector is $\underline{B}' = (\gamma + \alpha\underline{\underline{\mathrm{R}}} + \beta\underline{\underline{\mathrm{N}}}) \cdot \underline{B}$. This makes the overall measure

$$\frac{\underline{A} \cdot (\gamma + \alpha\underline{\underline{\mathrm{R}}} + \beta\underline{\underline{\mathrm{N}}}) \cdot \underline{B}}{|\underline{A}| \left|\gamma + \alpha\underline{\underline{\mathrm{R}}} + \beta\underline{\underline{\mathrm{N}}}\right| |\underline{B}|}.$$

The formula is symmetric: it is the same as the cosine measure between $\underline{B}$ and $\underline{A}' = (\gamma + \alpha\underline{\underline{\mathrm{R}}} + \beta\underline{\underline{\mathrm{N}}}) \cdot \underline{A}$ because the relationship matrices are symmetric.

3.1 Determining the Coefficients

To determine suitable values for the coefficients α and β, we choose them to maximize the overall consistency of professional human indexers. We take the work of human indexers to be the gold standard, and take thesaurus relations into account in a way that optimizes their performance.

Given terms assigned by a group of indexers to a group of documents, calculate the similarity between each indexer and all the others taken together, summed over all documents. This measures the degree to which that indexer agrees with the rest. Then choose α and β to maximize the total of these agreement values, in other words, maximize

$$SIM = \sum_{\text{indexers } i} \sum_{\text{documents } D} \frac{I_i^D \cdot (\gamma + \alpha\underline{\underline{\mathrm{R}}} + \beta\underline{\underline{\mathrm{N}}}) \cdot \sum_{\text{indexers } j} I_j^D}{\left|I_i^D\right| \left|\gamma + \alpha\underline{\underline{\mathrm{R}}} + \beta\underline{\underline{\mathrm{N}}}\right| \left|\sum_{\text{indexers } j} I_j^D\right|}$$

where I_i^D is the vector of terms that indexer I assigns to document D (and $\gamma + \alpha + \beta = 1$).

We gradually increased values for α and β to find the optimal values in the joint distribution for all indexers, which resulted in $\alpha = 0.20$ and $\beta = 0.15$ respectively. The peaks in the plotted data were shallow, which indicates that these values are approximate. The optimal values of α and β for the 5-indexer subsets range throughout the intervals [0.2,0.25] and [0.15,0.24] respectively.

When computing the cosine measure between indexers using the best overall values, if a term is the same in both sets it counts with a weight of 65%; if a term in one set is RT with a term in the other it counts with a weight of 20%; if it is BT/NT with the other term it counts with a weight of 15%. If it is not related to any term in the other set, its weight is 0. This simple and intuitive interpretation of weights demonstrates the advantage of working with the vector space model.

The re-analysis of our data with this new measure has shown that virtually all figures increase over the original version with $\alpha = \beta = 0$, although the overall cosine measure was very similar. Of course, our data set is certainly not large enough to judge indexers' performance, and our parameter estimation method—optimizing the overall performance of this particular set of indexers—is not conducive to highlighting differences between them.

4. SUMMARY

Existing measures of indexing consistency are flawed because they ignore semantic relations between the terms that different indexers assign. This paper has shown how the vector space model that underlies the cosine metric supports an elegant linear generalization of similarity that takes thesaurus relations into account. We introduce coefficients that reflect the relative importance of the thesaurus relations to the term-identity relation. We choose their values to optimise the performance of a set of professional human indexers. Alternatively, for request-oriented indexing, where a document's retrievability is more important than the consistency of its representation, the weights could be derived from searchers' relevance judgements.

We plan to use this measure to assess the quality of automatically produced keyphrases and to compare them with ones extracted by human indexers. Analysis of the conceptual relations between the phrases instead of simple matching of their stems will provide a sounder basis for judging the usability of automatic extraction in real-world applications.

5. ACKNOWLEDGMENTS

We gratefully acknowledge Jim Weinheimer and his team of indexers at the FAO for providing us with experimental data, and for enlightening discussions on the nature of the indexing task; and Dagobert Soergel for helping us get into the literature on this topic and for his critical assistance with drafts of this paper.

6. REFERENCES

[1] Hooper, R.S. (1965). *Indexer consistency tests—Origin, measurements, results and utilization.* IBM, Bethesda.

[2] Iivonen, M. (1995). Consistency in the selection of search concepts and search terms. *Information Processing and Management*, 31.

[3] Markey, K. (1984). Inter-indexer consistency tests. *Library and Information Science Research*, 6, 155–177.

[4] Rolling, L. (1981). Indexing consistency, quality and efficiency. *Information Processing and Management*, 17, 69–76.

[5] Zunde, P., & Dexter, M.E. (1969). Indexing consistency and quality. *American Documentation*, 20, 259–26