Dietmar Wolfram & Hope A. Olson
School of Information Studies, University of Wisconsin-Milwaukee
Milwaukee, Wisconsin, U.S.A. 53201

A Method for Comparing Large Scale Inter-indexer Consistency Using IR Modeling

Abstract: The authors present a method for comparing indexing consistency between groups of indexers based on the vector space IR model. Terms assigned by indexers are treated as vectors whose distances from a central vector may be compared. The method is outlined and demonstrated with an example.

Résumé : Les auteurs présentent un modèle pour comparer la cohérence interindexeurs entre des groupes d'indexeurs basé sur le modèle de RI d'espace vectoriel. Les termes attribués par les indexeurs sont traités comme des vecteurs avec lesquels il est possible de comparer la distance par rapport à un vecteur central. La méthode est expliquée et illustrée avec un exemple.

1. Introduction

The goal of indexing is to select and name topics that identify the aboutness of a document. The indexing terms selected by an indexer should also be identifiable by potential information searchers, not to mention other indexers. Ideally, if two indexers were to index the same document, each should identify the same or similar topics to indicate the aboutness of the document. The level of agreement provides an indication of inter-indexer consistency. Decades of research on consistency between indexers and by the same indexer at different times has documented medium to high levels of inconsistency. People simply do not choose the same concepts or the same words for those concepts. Consistency in indexing has long been considered essential for effective retrieval.

A number of measures for ascertaining consistency have been developed. To date, most of these measures can only be used to compare consistency between two indexers. Traditionally, this has been sufficient. Indexing, by nature, has been a solitary activity. Most documents are not usually indexed by more than one or two indexers, or if an information searcher identifies potential search terms, these are compared to those terms assigned to the document by an indexer. The need for measures of indexing that go beyond comparing two indexers was not apparent. The activity of indexing, however, has recently gained much wider appeal. The concept of "tagging" is essentially indexing performed by anyone interested in sharing their identified terms to describe a document in a public space (usually World Wide Web resources that provide a tagging feature). Today's tagging environments (e.g., Flickr, deli.icio.us, Amazon) allow many individuals to provide index terms for publicly available content to provide additional access points to these documents, whether they represent text, images, or digitized video. The current phenomenon of collaborative or social tagging changes the landscape of consistency research by adding the element of mass end-user indexing and permits the study of regularities in the way information is produced. Traditional methods for assessing inter-indexer consistency become inappropriate for these environments. This research presents a novel way of assessing inter-indexer consistency on a larger scale by relying on established information retrieval (IR) theory.

2. Previous Research

2.1 Inter-indexer Consistency

Inter-indexer consistency studies all point to high levels of inconsistency. Foundational studies on this topic go back to the 1960s, exemplified by the research undertaken by Zunde and Dexter (1969), whose analysis of earlier data (Schultz, Schultz, & Orr, 1965) pointed toward "power laws" and fuzzy sets as explanatory devices. Cooper's (1969) contemporary study questioned of the importance of consistency. Later studies were variations on these themes, branching out to other contexts. Markey (1984) performed a meta-analysis of 25 studies primarily from the 1960s and 1970s and noted not only the ubiquity of inconsistency from levels of 82% consistency down to 4%, but also looked at factors such as exhaustivity (i.e., number of terms assigned to a document) and vocabulary size that influence those levels. Markey also explored consistency in the indexing of images. Other studies also examined different information access contexts. Chan (1989) tested consistency in a library catalog context, refining definitions of a match and partial match to suit a precoordinate vocabulary, with similar results. Reich and Biever (1991), Sievert and Andrews (1991), Giral and Taylor (1993) and Leininger (2000) looked at indexing in agriculture, information science, and psychology respectively. These four studies, as well as those by Chan (1989) and Tonta (1991) were motivated by the existence of duplicate records either within or between databases. The authors examined terms assigned to two records for each of multiple titles. More recently, Saarti (2001) tested consistency in indexing and searching for fiction using a controlled vocabulary. Other perspectives also have been taken. For example, David et al (1995) explored the dynamics of inconsistency as a cognitive issue. Further examples generally follow the same patterns of investigation, investigating either multiple records from a serendipitous duplication or experiments involving multiple indexers or searchers and few records.

Looking in a different direction, Zunde and Dexter (1969) identify the implication of their data that a power law is operating in group indexing, suggesting that further informetric exploration could be fruitful. Maron's (1977) concept of retrieval aboutness or *R-about*, suggests that the ideal index terms are those that a given population who would find a document useful would use to search for that document. Maron was calling for a consensus that is obviously not found in previous inter-indexer consistency studies.

Several quantitative assessments of inter-indexer consistency have been developed to provide objective measures by which the level of consistency between two (or a few) indexers may be ascertained. Two frequently applied measures, as outlined by Medelyan and Witten (2006), include measures by Hooper (1965) and Rolling (1981). They rely on simple formulaic representations of the set terms assigned by each indexer and the common terms between the two sets. Hooper's measure defines the level of consistency between two indexers as the total number of terms in agreement divided by the total number of distinct terms used by both indexers (essentially the "anding" of both sets divided by the "oring" of both sets). Rolling's measure uses a variation of this approach. Medelyan and Witten demonstrate how the cosine measure, traditionally used in vector-based information retrieval to determine document proximities, could be used to calculate the level of consistency between two indexers.

2.2 Collaborative Tagging as Large-scale Indexing

The phenomenon of collaborative or social tagging is, in essence, indexing by non-professionals without the benefit of a controlled vocabulary. In each case users supply tags (terms) to describe individual items in a shared database such as bookmarked web

pages (http://del.icio.us/), uploaded photographs (http://flickr.com/), or books or other media (http://www.amazon.com). Some sites display terms that other users have used for the particular item. Collectively the tags are sometimes referred to as folksonomies and may display some hierarchy (typically very shallow). Collaborative tagging has become popular as one of the growing examples of social software now available on the WWW.

Golder & Huberman (2006) point to both positive and negative possibilities for collaborative tagging noting its "potential to exacerbate the problems associated with the fuzziness of linguistic and cognitive boundaries" (201). However, they also observe that tagging provides a learning opportunity for better understanding the sharing and organization of information. They examined data from del.icio.us for both tags and users to create a picture of the activity on the site. Their data indicated some tendencies that suggest that the process of tagging parallels indexing in a number of characteristics. Therefore, what we know about indexing may well be applicable to tagging. However, Golder and Huberman did not track consistency. In contrast, Kipp and Campbell (2006) did examine frequency distributions and coocurrences in del.icio.us. They found some parallels with indexing practice and concluded that there are similiarities in the tendency toward a core of terms widely used by different taggers/indexers. Further, they propose the possibility that clusters of synonyms may give broader access than the mutually exclusive headings of controlled vocabularies. The user-defined nature of the tags allows for functions that go beyond representing aboutness to practical applications.

3. Defining Indexing Spaces Generated by Groups of Indexers

Existing measures of inter-indexer consistency are limited in that they only allow up to several (usually two) indexers to be compared. With today's social environment, where potentially hundreds of people may index the same document, these measures cannot be used to compare indexing outcomes by groups of indexers. Medelyan & Witten propose a new measure based on thesaurus relationships of assigned terms by treating term assignments using the familiar vector space model found in information retrieval. The authors' method assumes the availability of a thesaurus through which semantic relationships between assigned terms may be determined.

In environments where a controlled vocabulary is not used, this approach would not work. However, the use of the vector space model to represent indexing spaces among large numbers of indexers is still feasible. Briefly, the terms assigned by a group of indexers to a document may be thought of as similar to the concept of a document space used in information retrieval. In this classic vector space model developed by Salton (1975), documents are represented as vectors in an *n*-dimensional space. The closeness of the relationship between documents is based on the calculated similarity between documents. A document matrix *V* for a document set consisting of *m* documents and *n* terms will take the form:

$$V = \begin{bmatrix} t_{11} & t_{12} & \dots & t_{1n} \\ t_{21} & t_{22} & \dots & t_{2n} \\ \dots & \dots & \dots & \dots \\ t_{m1} & t_{m2} & \dots & t_{mn} \end{bmatrix}$$

where t_{ij} represents the *i*th term in the *j*th document, representing a high-dimensional document space. If term *i* exists in document *j*, its value is set to 1 or an appropriate weighted value for its significance; otherwise it is set to 0. Because of the potentially large numbers of documents and the large numbers of terms, most document vector elements will be 0.

This same concept can be applied to indexing environments, where instead of documents, rows of the matrix are represented by indexers' choices for term selection for a given document. With a representation method for large numbers of indexers, a method of comparing groups of indexers is needed. Measures of closeness between documents and queries or documents themselves have been developed in classic IR. They measure distances or angles between pairs of documents, but do not provide a means of measuring overall document space characteristics. A measure that has been developed to provide an indication of the cohesiveness of a document space is that of a document space density (Salton, 1975). This has also been applied by Wolfram & Zhang (2001, 2002) to measure the effect of the addition and removal of one or more index terms from a document set on the overall document space. In IR, the more closely, or densely, documents are situated to one another, the more indistinguishable they are, which may negatively affect retrieval outcomes. Therefore for more effective IR, a diffuse or less dense document space is desirable for IR. For indexing consistency purposes, the opposite is true, where a dense indexing space is more indicative of similarity and higher levels of consistency.

Indexing consistency comparisons across groups of indexers may be conducted by first calculating the distance between each indexer's resulting vector and the indexing centroid (or average vector across all indexers). Figure 1 demonstrates this concept.

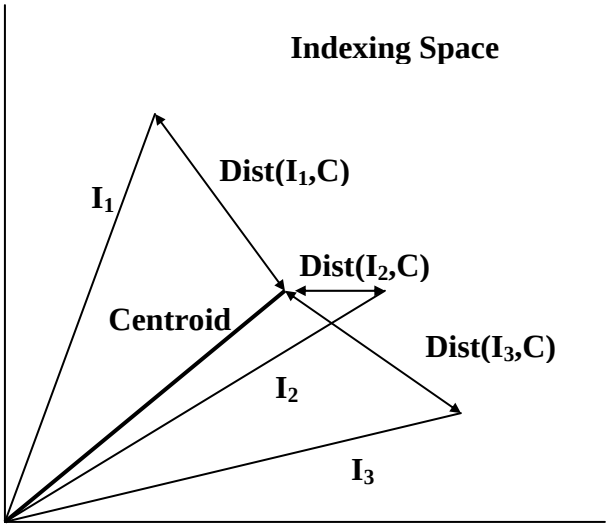

Figure 1. Indexer Distances from the Indexing Centroid

The average distance of these values provides an indication of the Inter-indexer Consistency Density (ICD). For a group of *m* indexers (*I*) who index a given document:

$$ICD = \frac{\sum_{i=1}^{m} Sim(I_i, C)}{m} \quad (1)$$

$$\text{where } Sim(I_i, C) = \begin{cases} \dfrac{1}{Dist(I_i, C)} & for\ Dist(I_i, C) \neq 0 \\ 1 & for\ Dist(I_i, C) = 0 \end{cases}$$

where $Dist(I_i, C)$ represents the Euclidean distance between I_i and centroid *C*. The centroid is an *n*-dimensional vector where *n* is the number of distinct terms in the indexing space. The value assigned to each element of the centroid is the total number of occurrences of the specific term (i.e., its frequency) divided by the total number of indexers. The density measure, which represents the mean distance of indexer vectors from the centroid, serves as a normalized method for comparing the indexing space characteristics of each environment studied.

To compare if significant differences arise between or among groups, each similarity value ($Sim(I_i, C)$) can serve as an observation for which appropriate statistical tests may be run. In the case of two groups of indexers, a t-test (or non-parametric Mann-Whitney test) could be used to compare the ICDs for each group. For more than two groups, a one-way Analysis of Variance test (or non-parametric Kruskal-Wallis test) could be used. Note that this method may also be applied to the same group of indexers across different documents to assess group consistency over time.

4. An Example

To demonstrate an example of the method at work, two indexing spaces were derived from responses by MLIS students. Data were collected from four sections (two online, two onsite) of a course on Information Organization, over two semesters. Students were asked to provide up to five terms that describe what an assigned conference paper was about. Aggregate data were stored and tabulated in a MS Access database based on the frequency of occurrence of terms assigned. Terms were not from a controlled vocabulary and minimal regularization (e.g. collapsing singular and plural occurrences) was employed. Data collected represent two semesters of student indexing data. The data sets comprise the selected index terms identified by 33 (Group 1) and 31 students (Group 2), respectively, each representing different semesters of students. Additional detail regarding indexing analysis of this data may be found in Olson and Wolfram (submitted).

A routine was written to calculate similarity measures for each indexer to the centroid for each group. Resulting values were then used to run t-tests in SPSS on the similarity values.

	Mean (ICD)	Standard Deviation
Group 1	0.4398071	0.0939877
Group 2	0.4275383	0.0220765

t-test outcome (assuming unequal variances)
$t = 0.7288 \quad \alpha = .05 \quad p = 0.471$

Table 1. Indexing Consistency Comparison Outcome

In this case, the level of consistency between the two groups was not significantly different.

5. Discussion & Conclusions

The question arises: so what? How does the ability to compare consistency among groups of indexers inform indexing practice or ultimately improve retrieval? As the process of social indexing becomes more ubiquitous, the ability to assess changes in indexing behavior among groups of indexers or within groups of indexers over time becomes possible. A multiple indexer comparison method also allows comparisons to be made across languages or cultures. In educating users about indexing practice, a method for measuring difference also allows the effectiveness of indexing training to be assessed, particularly if controlled vocabularies with more limited options are used.

A limitation of this assessment approach is that consistency measures are based on the similarity of term usage based on their frequencies, which is independent of term meaning. Differences are based solely on quantitative assessments of distance in the indexing space. Indexers who assign the same number of terms to a document that occur with the same frequency will result in the same distance from the centroid. The frequency distribution of index terms assigned shows a strong inverse pattern that is Zipf-like (Golder & Huberman, 2006; Olson & Wolfram, submitted), where many index terms occur only once, raising the possibility of similar distances. However, with variability in indexing exhaustivity, this likelihood is reduced. Also, the present example assumed binary assignment of term weights, where a term was either present (weight = 1.0) or absent (weight = 0.0). More fine-grained assignment of weights may also be used to distinguish term importance, but this is not frequently done in indexing practice beyond indicating whether a term is assigned major or minor status.

Relying on a vector space model makes it possible to compare indexing consistency using many indexers or groups of indexers. Future research will examine the consistency in public environments, such as social tagging sites, where indexing practices of different groups of indexers may be compared or the same indexers are compared across different documents or over time.

Acknowledgements

We are indebted to our colleague Jin Zhang for his insights into vector space modelling as well as his feedback.

References

Chan, L.M. 1989. Interindexer consistency in subject cataloging. *Information Technology and Libraries* 8: 349-358.

David, C., et al. 1995. Indexing as problem solving: A cognitive approach to consistency. In *Proceedings of the ASIS Annual Meeting* 32: 49-55.

Giral, A., & A.G. Taylor. 1993. Indexing overlap and consistency between the Avery Index to Architectural Periodicals and the Architectural Periodicals Index. *Library Resources & Technical Services* 37: 19-44.

Golder, S.A., & B.A. Huberman. 2006. Usage patterns of collaborative tagging systems, *Journal of Information Science* 32: 198-208. [an apparently earlier version is available under the title: The structure of collaborative tagging systems at http://www.hpl.hp.com/research/idl/papers/tags/tags.pdf].

Hooper, R. S. 1965. *Indexer consistency tests—Origin, measurements, results and utilization*. Bethesda: IBM.

Kipp, M.E.I, & D.G. Campbell. 2006. Patterns and inconsistencies in collaborative tagging systems: An examination of tagging practices. In *Proceedings of the ASIST Annual Meeting* 43: [CD-ROM].

Leininger, K. 2000. Interindexer consistency in PsycINFO. *Journal of Librarianship and Information Science*. 32: 4-8.

Markey, K. 1984. Interindexer consistency tests: A literature review and report of a test of consistency in indexing visual materials. *Library & Information Science Research* 6: 155-177.

Maron, M.E. 1977. On indexing, retrieval and the meaning of "about". *Journal of the American Society for Information Science* 28: 38-43.

Medelyan, O., & I. Witten. 2006. Measuring inter-indexer consistency using a thesaurus. In *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*. pp. 274-275. New York: ACM Press.

Olson, H.A. & D. Wolfram. Submitted. Syntagmatic relationships and indexing consistency on a larger scale.

Reich, P., & E.J. Biever. 1991. Indexing consistency: The input/output function of thesauri. *College & Research Libraries* 52: 336-342.

Rolling, L. 1981. Indexing consistency, quality and efficiency. *Information Processing & Management* 17: 69-76.

Saarti, J. 2001. Consistency of subject indexing of novels by public library professionals and patrons. *Journal of Documentation* 58: 49-65.

Salton, G. 1975. *A theory of indexing*. Philadelphia: Society for Industrial and Applied Mathematics.

Schultz, C.K., W.L. Schultz, & R.H. Orr. 1965. Comparative indexing: Terms supplied by biomedical authors and by document titles. *American Documentation* 16: 299-312.

Sievert, M.C., & M.J. Andrews. 1991. Indexing consistency in Information Science Abstracts. *Journal of the American Society for Information Science* 42: 1-6.

Tonta, Y. 1991. A study of indexing consistency between Library of Congress and British Library catalogers. *Library Resources & Technical Services* 35: 177-185.

Wolfram, D., & J. Zhang. 2001. The impact of term indexing characteristics on a document space. *Canadian Journal of Information and Library Science* 26: 21-35.

Wolfram, D., & J. Zhang. 2002. An investigation of the influence of indexing exhaustivity and term distributions on a document space. *Journal of the American Society for Information Science and Technology* 53: 943-952.

Zunde, P., & M.E. Dexter. 1969. Indexing consistency and quality. *American Documentation* 20: 259-267.