Evaluating Publication Similarity Measures

Sulieman Bani-Ahmad, Ali Cakmak, and Gultekin Ozsoyoglu
EECS Dept,
Case Western Reserve University
(sulieman, ali.cakmak, tekin) @case.edu

Abdullah Al-Hamdani
Dept. of Computer Science
Sultan Qaboos University, Oman
abd@squ.edu.om

Abstract

Publication searching based on keywords provided by users is traditional in digital libraries. While useful in many circumstances, the success of locating related publications via keyword-based searching paradigm is influenced by how users choose their keywords. Example-based searching, where user provides an example publication to locate similar publications, is also becoming commonplace in digital libraries.

Existing publication similarity measures, needed for example-based searching, fall into two classes, namely, text-based similarity measures from Information Retrieval, and citation-based similarity measures based on bibliographic coupling and/or co-citation.

In this paper, we list a number of publication similarity measures, and extend and evaluate them in terms of their accuracy, separability, and independence. For evaluation, we use the ACM SIGMOD Anthology, a digital library of about 15,000 publications.

1 Introduction

Searching publications based on keywords is common in digital libraries. While useful in many circumstances, the success of locating related publications based on keywords depends on the choice of keywords [6]. Example-based searching, i.e., locating similar/related publications to a given publication is also becoming a common search query type in digital libraries [13]. In this work, we deal with the quality of publication similarity measures used for locating related- or similar-publications of a given publication. Existing publication similarity measures fall into two classes: (i) text-based similarity measures from the field of Information Retrieval (IR), such as the cosine similarity and the TF-IDF (term frequency-inverse domain frequency) model [14], or (ii) citation-based similarity measures based on bibliographic coupling (i.e., common citations between two publications) [8], co-citation (i.e., common citers of two publications) [15] or author-coupling (i.e., common authors between two publications). In this paper, we summarize the existing publication similarity measures, and extend and evaluate them in terms of their *accuracy*, *separability*, and *independence*. For evaluation, we use the

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

ACM SIGMOD Anthology [1], referred to as *AnthP* here, a digital library of about 15,000 publications in data management.

Text-based similarity measures are based on information retrieval methodologies [14, 5]. As an example, using the vector space model of IR and the TF-IDF weighting scheme [14], the similarity between two publications may be measured by using Cosine, Jaccard, Dice or other document measures [10].

CiteSeer [2] is a literature search system for searching (presently) about 730,000 computer science and bioinformatics publications, and uses three document similarity measures, namely, word vectors, LikeIt string distance, and the Common Citation Inverse Document Frequency [7]. Google Scholar, Google scholarly literature search engine [3], does not provide publication similarity functions which are needed to answer example-based queries where the user provides an example publication and asks for similar publications.

By evaluating "multiple levels" of paper similarities based on bibliographic-coupling, co-citation and author-coupling, we make the following observations:
(a) Similarity value distribution curves are similar within the same group of similarity measures, i.e., bibliographic-coupling-based, co-citation-based, and author-coupling-based measures,
(b) Citation-based and author-coupling-based similarity measures are more separable than bibliographic-coupling-based measures,
(c) Citation-based and author-coupling-based similarity measures are all highly correlated. This phenomena is due to the citation and coauthorship behavior in the literature [11].
(d) Text-based similarity measures show low overlapping with citation-based and with author-coupling-based measures. Therefore, providing two sets of similarity scores, one text-based and another based on citation and/or author-coupling may prove to be a useful practice.

This paper is organized as follows. In section 2, we list and extend a number of publication similarity measures. In section 3, we evaluate the proposed similarity measures. Section 4 concludes.

2 Similarity Measures between Two Publications

2.1 Text-Based Similarities

The vector space model of text documents is used to evaluate title, abstract, index terms, and body similarities between two papers [14]. Consider a vocabulary T of atomic terms t that appear in each document. A document is represented as a vector of real numbers $v \in R^{|T|}$, where each element corresponds to a term. Let v_t denote an element of v that corresponds to the term t, $t \in T$. The value of v_t is related to the importance of t in the document represented by v. Using the Term Frequency-Inverse Document Frequency (TF-IDF) weighting scheme [14], v_t is defined as

$$v_t = log(TF_{v,t} + 1) * log(IDF_t)$$

where $TF_{v,t}$ is the number of times that term t occurs in the document represented by v, $IDF_t = N/n_t$, N is the total number of documents in the database, and n_t is the total number of documents in the database that contain the term t.

The cosine similarity between two documents with vectors v and w is computed as

$$cosine(v, w) = (\sum_{i=1}^{|T|} f(v_i).f(w_i))/(\sum_{i=1}^{|T|} f(v_i)^2.\sum_{i=1}^{|T|} f(w_i)^2)$$

where f() is a damping function, which is either the square-root or the logarithm function. Other similarity functions include Dice and Jaccard measures [10] where both change the normalization factor in the denominator to account for different characteristics of the data. As a preprocessing step, one needs to first remove the stopwords from the terms of a document, and then use the Porter's algorithm [12] to stem the terms.

2.2 Citation-Based Similarities

The citation-based similarity between two publications can be computed using (a) bibliographic coupling: common citations between the two publications [8], and (b) co-citation: common citers to the two publications [15]. One can then define citation-based similarity between two publications as a weighted sum of the two. In this section, we discuss various ways of computing bibliographic coupling and co-citation.

2.2.1 Bibliographic Coupling with Reachability Analysis

The bibliographic coupling-based similarity between papers P_Q and P_X, $Sim_{bib}(P_Q, P_X)$, can be defined as

$$Sim_{bib1}(P_Q, P_X) = \text{(common citations count between } P_Q \text{ and } P_X\text{)/MaxB.}$$

where MaxB is the maximum number of common citations between any two publications in $AnthP$. One problem with this definition is that it assumes that each common citation contributes to the reference similarity equally, and ignores the effects of publications that are cornerstone works leading to significant research in the field. A cornerstone publication is cited by all the publications that discuss an issue related to the field, and its citation by two publications carries a lesser significance. Hence it is quite possible for two publications about two unrelated topics to cite the same cornerstone publication.

To reduce the effect of common citations to cornerstone works, we define a new bibliographic coupling measure where each common citation contributes at a different level depending on the extent to which it is "influential". Assume that we assign importance scores to publications using the well-known PageRank algorithm [4]. PageRank scores are computed recursively using the formula $P_{i+1} = (1-d)M^T P_i + E$ where P_{i+1} and P_i are the current and next iteration PageRank vectors respectively, citation matrix C is the adjacency matrix of a graph with papers representing nodes, and citation relationships between papers representing edges, M is a matrix derived from C by normalizing all row-sums in C to 1, and, d is the "future citation probability" defined as follows. Given (a) an author A writing a new paper and citing paper u which in turn cites paper v, and (b) w, a randomly selected paper in $AnthP$, the parameter d, which we choose to be low, represents the probability that A will cite w, and $(1-d)$ is the probability that A will cite v. C is of size $N \times N$, where N is the total number of papers in the system. To guarantee that the PageRank algorithm converges, a hidden link, represented by the user-defined parameter E, is assumed to exist between each pair of graph nodes. A choice for E is simply $E_1 = d$. Another choice, used in [4], is $E_2 = d/N[1_N].P_i$ where 1_N is a vector of N ones. A highly important publication is cited by a large set of publications, and therefore, cannot provide an informative measure. On the other hand, if two publications cite a publication with a relatively low importance score, this citation information provides more clues toward the similarity of the two publications. Therefore, we assign weights to common citations, which are inversely proportional to their (importance) scores as follows.

$$Sim_{bib2-L1}(P_Q, P_X) = \textstyle\sum_{P_i \in S_{QX}} (1 - P_{Score}(P_i))/MaxW$$

where S_{QX} is the set of common citations between P_Q and P_X, $P_{Score}(P_i)$ is the PageRank-based score of paper P_i. MaxW is the maximum $\sum_{P_i \in S_{QX}} (1 - P_{Score}(P_i))$ for any two publications in ***AnthP***.

Another extension to bibliographic coupling similarity is to incorporate the notion of citations iteratively, which we refer to as *reachability analysis*. The formula of $Sim_{bib2-L1}$ can be considered as the $firstlevel$ (level-1) evaluation of a given citation information. We can also make use of $second-level$ and $third-level$ citation information. Due to efficiency considerations, next we consider only the most basic reachability analysis cases. Normally if a publication is cited by only one of the publications (i.e., either P_Q or P_X, but not both) then this publication is not considered in $Sim_{bib2-L1}$. Nevertheless, by following the citation information one more level, we may obtain additional information. For instance, assume that publication P_i is cited by P_Q, but not cited by P_X. It is possible that, at one level below, P_i may be cited by one of the publications, say P_j, which is in turn cited by P_X, as illustrated in Figure 1(a).

Note that second-level common citations can be used to strengthen common citation information of publications P_Q and P_X. Assume that P_i is cited by both P_Q and P_X. This common citation may lead to more similarity clues such that P_i might cite a publication P_k which is cited by P_Q, P_X or both, as illustrated in Figure 1(b). Finally, third level common citations can be considered as common citations for publications P_Q and P_X which is illustrated in Figure 1(c).

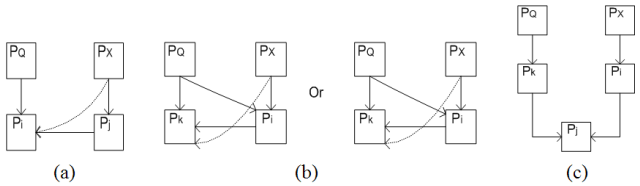

Figure 1: Illustration of citation networks (a) one level (b) two levels (c) three levels

We do not consider higher levels of co-citation information since, at each new level, publications get more diverse in terms of their contents, and their citations become less significant.

2.2.2 Co-citation Similarity with Reachability Analysis

As in multi-level bibliographic coupling, we can apply the same one, two, or three-level co-citation similarity in a similar manner. Different co-citation cases are illustrated in Figure 2, and the corresponding co-citation definitions are given next. One-level co-citation similarity between papers P_Q and P_X is defined as

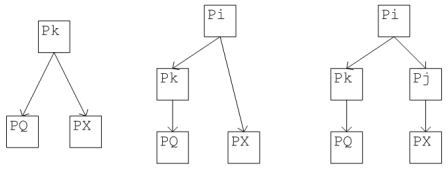

Figure 2: Illustration of three levels of co-citation similarity.

$$Sim_{co-cit1} = |C_Q \cup C_X|/MaxN$$

where C_Q, C_X are the set of publications each of which cites P_Q and P_X, respectively and $MaxN$ is the maximum number of common citers between any pair of publications in the $AnthP$. Once again, assume that we use a paper scoring algorithm, such as PageRank, to assign importance scores to publications. Then, if a publication citing P_Q or P_X is a hub (e.g., a survey paper) [9], then it will refer to many publications. To reduce the effects of hubs, we use

$$Sim_{co-cit2-L1} = \sum_{P_i \in S_{QX}} (1 - P_{Score}(P_i))/MaxC$$

where S_{QX} is the set of publications that co-cite P_Q and P_X, $P_{Score}(Pi)$ is the importance score of co-citer P_i, MaxC is the maximum $\sum_{P_i \in S_{QX}} (1 - P_{Score}(P_i))$ value of any pair of publications in $AnthP$.

If publications P_Q and P_X are cited together by more than one publication, then we can weigh the contribution of each citing publication by its "hub score" of HITS [9]. Here we use the hub score of the citing publication because this relationship represents an outgoing link from the citing publication to P_Q and P_X. For outgoing links, in Kleinberg's model [9], the hub score of the entity determines the strength of the outgoing link. Therefore if the citing publication is a good hub with a relatively high hub score then it contributes more than other citing publications rather than each citing publication contributing equally. Thus, we have yet another co-citation-based function:

$$Sim_{co-cit-Hub} = (\sum_{P_i \in S_{QX}} (1 - P_{HubScore}(P_i)))/MaxCh$$

where $P_{HubScore}(P_i)$ is the hub score of publication P_i, and $MaxCh$ is the maximum $\sum_{P_i \in S_{QX}} (1 - P_{HubScore}(P_i))$ value between any pair of publications in *AnthP*.

2.3 Author-Coupling-Based Similarities

We compute the *author similarity* between two publications directly via the number of common authors between the two publications (referred to as the Level-0-author overlap Sim_{AOC-L0}) or indirectly via co-authorship in other publications, e.g., two different authors, each of different publications P_Q and P_X, are co-authors in a third publication P_W (referred to here as the Level-1-author-overlap Sim_{AOC-L1}). We then use the following formula to compute the author similarity between publications P_Q and P_X:

$$Sim_{Author}(P_Q, P_X) = W_{L0} * Sim_{AOC-L0}(P_Q, P_X) + (1 - W_{L0}) * Sim_{AOC-L1}(P_Q, P_X)$$

where $0 \leq W_{L0} \leq 1$ and

$$Sim_{AOC-L0}(P_Q, P_X) = |A_Q \cup A_X|/MaxA0$$
$$Sim_{AOC-L1}(P_Q, P_X) = (1/MaxA1) \sum_{(i \in A_Q) \wedge (j \in A_X)} |(S_i - \{P_Q\}) \cup (S_j - \{P_X\})|$$

where A_Q and A_X are the sets of authors of P_Q and P_X, respectively. S_i and S_j each is the set of papers written by authors i and j, respectively, where $i \in A_Q$ and $j \in A_X$. $MaxA0$ and $MaxA1$ are the maximum numbers of level 0 (L_0) and level 1 (L_1) co-author overlaps, respectively, of any two publications in *AnthP*.

Next we assume that we have importance scores computed for authors. As an example, we may compute an author importance score as the average of importance scores assigned to the author's perhaps top-k publications. Then, as another variant, we can also consider using a different mechanism so that each shared author contributes to the similarity of publications in different proportions, depending on his/her author importance scores. This is based on the assumption that the works of important authors share a common thread. As an example, we produce a higher similarity score for two publications which share one author with a high importance score in comparison with two publications which share one author with a low ranking. On the other hand, in practice, with some exceptions, well-known authors are usually the ones who publish many high quality publications. Moreover, due to their prolificacy, it is not uncommon for these authors to publish on relatively different topics. Therefore we use a weighing mechanism which leads to author weights that are inversely proportional to their importance scores. In this way, the information that two publications share a less important author implies more towards the similarity of the publications in comparison to the case that these publications share an author with a higher importance score. Thus, we define the Level-0 and level-1 author-overlap involving author weighting Sim_{AOW-L0} and Sim_{AOW-L1} as follows

$$Sim_{AOW-L0}(P_Q, P_X) = \sum_{a_i \in A_{QX}} (1 - A_{Score}(a_i))/MaxA0$$
$$Sim_{AOW-L1}(P_Q, P_X) =$$
$$(1/MaxA1) \sum_{(i \in A_Q) \wedge (j \in A_X)} (1 - A_{Score}(a_i))(1 - A_{Score}(a_j))|(S_i - \{P_Q\}) \cup (S_j - \{P_X\})|$$

where A_Q and A_X are the sets of authors of publications P_Q and P_X, respectively, A_{QX} is the set of common authors between P_Q and P_X. $MaxA0$ and $MaxA1$ are the maximum numbers of level 0 (L_0) and level 1 (L_1) co-author overlaps, respectively, of any two publications in *AnthP*. In our experiments, we compute the score $A_{Score}(a)$ of author a as the average score of most important K papers of a, where K is 5.

3 Empirical Evaluation of Publication Similarity Measures

3.1 Experimental Setup

For each publication in $AnthP$, we extracted titles, authors, publication venues, publication year information, and citations. The final experimental dataset included (a) 106 conferences, journals, and books, (b) 14,891

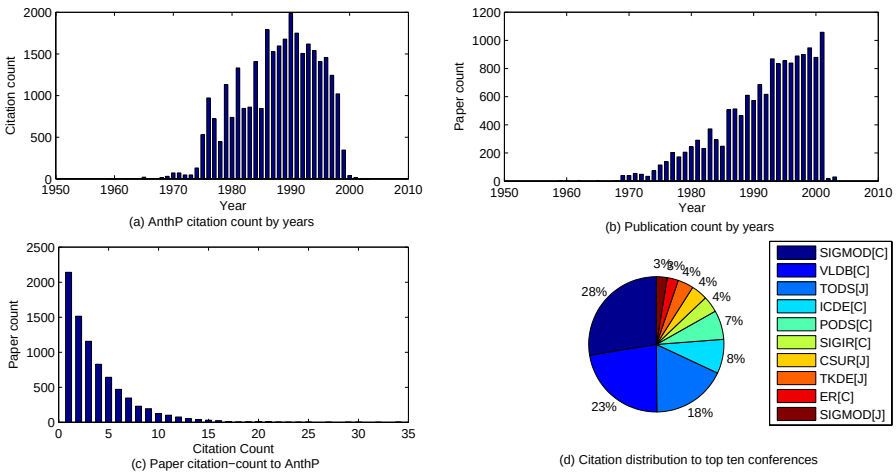

Figure 3: *AnthP* statistics

publications, and (c) 13,208 authors. *AnthP citation* refers to a citation from any publication in the *AnthP* set to a publication in the same set. *DBLP citation* refers to a citation from a publication in *AnthP* to a publication P outside of *AnthP*, but within DBLP. *External citation* of publication P is a citation from publication P to a publication outside of *AnthP* and *DBLP*.

Next we present $AnthP$ statistics. The average number of citations in an *AnthP* publication is 20. The average number of *AnthP* and DBLP citations in an *AnthP* publication is 4.289. The average *AnthP* citation count per *AnthP* publication is 2.066. Thus, the average citation reduction due to DBLP citation removal is 48.2%. Figure 3(a) displays the citation count distribution of *AnthP* publications over years, Notice that the most recent publications are not cited yet, which means that their scores will be very low even though we do not know how important they are for sure. Same comments apply to the publications published before 1974; we do not have information as to which publications cite them. The publications published before 1974 and after 2000 are very few as shown in Figure 3(b). Figure 3(c) displays the distribution of *AnthP* citation counts for the publications in *AnthP*. Figure 3(d) shows top ten venues in term of citation counts. We think that all ten venues are known to be among the best in the computer science community.

In section 3.2, we compare publication similarity measures in terms of $separability$, $independence$ and $accuracy$. $Separability$ refers to having similarity scores that distribute to a large range reasonably well. To compare similarity measures in terms of $separability$, we use similarity score distribution plots. $Independence$ refers to similarity measures that are not (highly) correlated. We evaluate $independence$ using pairwise Top-K overlapping ratios. We define the Top-K Overlapping ratio between two measures m_1 and m_2 as:

$$TKO(m_1, m_2) = Average_{(\forall p \in \textit{AnthP})}[SS_1(p) \cap SS_2(p)]/min(|SS_1(p)|, |SS_1(p)|)$$

where $SS_1(p)$ and $SS_2(p)$ are the sets of K most-similar publications to publication p based on m1 and m2, respectively. For our experiments, we used K=50. We do not consider publications with zero similarity in the set of similar publications. $Accuracy$ refers to how accurate a similarity measure is. For $accuracy$, we compute the overlapping between text-based and citation-based similarity measures, i.e., we consider text-based measure (in this case, TF-IDF and Cosine similarity) as a benchmark to which we compare citation-based similarity measures.

3.2 Experimental Results

Observation: (Figure 4): Paper similarity measure distribution within the same group of similarity measures are similar, where the groups are defined as bibliographic-coupling-based, co-citation-based, and author-coupling-based.
Observation: (Figure 4): Citation-based and author-coupling based similarity measures are more separable than bibliographic-coupling-based measures
Observation: Paper overlapping ratio within bibliographical coupling-based similarity measures outputs to the

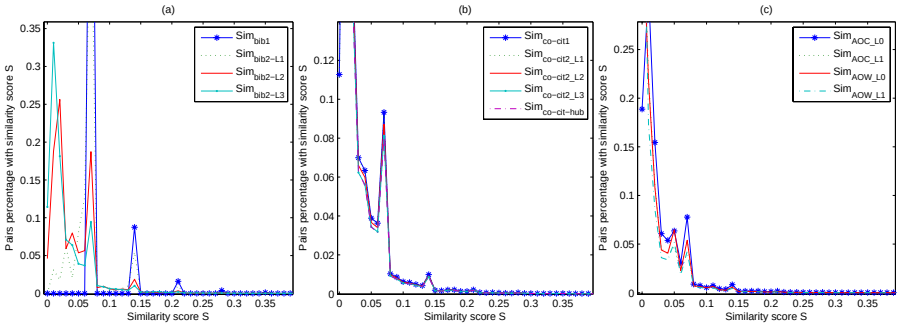

Figure 4: (a) Bibliographic-coupling-based, (b) citation-based and (c) author-coupling-based paper similarity score distributions.

same query ranges from 0.82 to .92.
The reason for the above observation is that, although a particular paper P usually deals with a limited and usually a single topic, its references cover a much wider range of research topics. This diversity increases by moving to the references of references. Thus,
Observation: In general, moving from a lower level to a higher level in bibliographical coupling-based measures creates more diversity, and in turn, smaller overlapping ratio.
Observation: Top-50 overlapping ratio between those similarity measure outputs based on bibliographical coupling and those based on co-citation ranges from 0.81 to 1.0.

The reason for the above observation is perhaps that authors usually tend to cite their own previous papers. On the other hand, most of one author's papers in general cover a small number of research interests which makes most of his/her work cite similar works. This leads to high top-50 overlapping paper ratios between the similarity measures based on bibliographical coupling and those based on co-citation.
Observation: Top-50 overlapping paper ratio between those similarity measures based on author-coupling overlapping and those based on co-citation ranges from 0.86 to 0.95.
Observation: Top-50 overlapping papers ratios between those similarity measures based on author-coupling and those based on bibliographical coupling ranges from 0.77 to 0.96.

The reason for the above observation is that, if two papers are similar based on an author-coupling measure then these papers in general are similar based on bibliographical coupling because the common authors usually have the same or at least somewhat related research interests. This makes the papers they publish commonly cite almost the same set of publications.
Observation: Text-based similarity measures show low overlapping with citation-based and author-coupling-based measures.
The above observation resulted from the way we retrieve top similar papers based on TF-IDF and Cosine similarity measure. That is, the papers that we find to be similar to a particular paper p are sorted according to their importance scores. Then we report top scored similar papers. This prevents papers that are similar, but low scored, to p also from appearing in the reported set. This in turn reduces the overlapping between text-based similarity measures in one side, and citation-based and author-coupling-based measures in the other side.

4 Conclusions

In this paper, we have presented and evaluated three groups of paper similarity measures in terms of their (i) *accuracy* (ii) *separability* and (iii) *independence*. For evaluation, we have used the ACM SIGMOD Anthology, a digital library of about 15,000 publications.

5 Acknowledgment

This research is supported by the US National Science Foundation grant ITR-0312200. S. Bani-Ahmad is supported by a fellowship from BAU-Jordan.

References

[1] ACM SIGMOD Anthology, http://www.acm.org/sigmod/dblp/db/anthology.html.

[2] CiteSeer Scientefic Digital Litrature Library, http://citeseer.ist.psu.edu/.

[3] Google Scholar (Beta), http://scholar.google.com/scholar/.

[4] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 1998.

[5] W. Cohen. The WHIRL approach to integration: An overview. In *Proceedings of the AAAI Workshop on AI and Information Integration*, Madison, Wisconsin, 1998.

[6] J. Dean and M. R. Henzinger. Finding related pages in the World Wide Web. *Computer Networks (Amsterdam, Netherlands: 1999)*, 31(11–16):1467–1479, 1999.

[7] L. Giles, K. D. Bollacker, and S. Lawrence. CiteSeer: An automatic citation indexing system. In *Proc. of Intl. Conf. Digital Libraries*, 1998.

[8] R. Johnson and D. Wichern. *Applied Multivariate Statistical Analysis*. Prentice Hall, Upper Saddle River, New Jersey, 1998.

[9] J. Kleinberg. Authoritative sources in hyperlinked environments. In *the 9th ACM-SIAM Symposium on Discrete Mathematics*, 1998.

[10] G. Kowalski. *Information retrieval systems: theory and implementation*. Kluwer Academic Publishers, 1997.

[11] M. Newman. Coauthorship networks and patterns of scientefic collaboration. *PNAS*, 2004.

[12] M. Porter. An algorithm for suffix stripping. *Program*, 14(3), 1980.

[13] C. S. Lawrence, L. Giles, and K. Bollacker. Digital libraries and autonomous citation indexing. In *Intl. Conf. Digital Libraries*, 1998.

[14] G. Salton. *Automatic Text Processing*. Addison Wesley, 1989.

[15] H. Small. Co-citation in the scientific literature: a new measure of the relationship between two documents. *Journal of the American Society for Information Sciences 24*, pages 265–269, 1973.