Research Paper

Mining Related Articles for Automatic Journal Cataloging

Yuqing Mao[1, 2†] & Zhiyong Lu[2]

[1]School of Information Technology, Nanjing University of Chinese Medicine, Nanjing 210023, China
[2]National Center for Biotechnology Information, National Library of Medicine, MD 20894, USA

Citation: Yuqing Mao & Zhiyong Lu (2016). Mining Related Articles for Automatic Journal Cataloging.

Received: Dec. 14, 2015
Revised: Jan. 5, 2016
Accepted: Feb. 26, 2016

Abstract

Purpose: This paper is an investigation of the effectiveness of the method of clustering biomedical journals through mining the content similarity of journal articles.

Design/methodology/approach: 3,265 journals in PubMed are analyzed based on article content similarity and Web usage, respectively. Comparisons of the two analysis approaches and a citation-based approach are given.

Findings: Our results suggest that article content similarity is useful for clustering biomedical journals, and the content-similarity-based journal clustering method is more robust and less subject to human factors compared with the usage-based approach and the citation-based approach.

Research limitations: Our paper currently focuses on clustering journals in the biomedical domain because there are a large volume of freely available resources such as PubMed and MeSH in this field. Further investigation is needed to improve this approach to fit journals in other domains.

Practical implications: Our results show that it is feasible to catalog biomedical journals by mining the article content similarity. This work is also significant in serving practical needs in research portfolio analysis.

Originality/value: To the best of our knowledge, we are among the first to report on clustering journals in the biomedical field through mining the article content similarity. This method can be integrated with existing approaches to create a new paradigm for future studies of journal clustering.

Keywords PubMed; Journals; Cluster; Catalog; Text mining; Research evaluation

† Corresponding author: Yuqing Mao (E-mail: mao.yuqing@msn.com).

JDIS
Journal of Data and Information Science
Vol. 1 No. 2, 2016
pp 45–59
DOI: 10.20309/jdis.201613
http://www.jdis.org

Research Paper

1 Introduction

In recent years, the research on clustering journals of similar topics has attracted much attention of scientists. One important reason is that it is an intermediate step towards research portfolio analysis, which is the analysis of research programs that can be classified by any theme of interest, including those related to administrative needs, organizational structure, funding streams, goals, and results (Srivastava, Towery, & Zuckerman, 2007). In such kind of analysis, journals need to be first grouped and sorted in the same category for assessing the significance of research in a specific field. Journal clustering is useful not only for classification, but also for indexing and retrieving schemes (Shultz, 2007; Small & Koenig, 1977). For example, the results of journal clustering can be utilized to index journals based on each journal's research area to improve journal search accuracy and efficiency.

Traditionally, journal clustering is based on human cataloging. However, this manual approach is unable to 1) keep up with the rapid growth of new journals, 2) capture the changes in journal scopes over time, and 3) measure the relatedness between journals. The number of scientific journals has rapidly increased over the last decades, especially in the active and productive biomedical area. Scientists are facing thousands of new journals in PubMed every year, as illustrated in Figure 1. In addition, the scope of a journal may change (Kang, Doornenbal, & Schijvenaars, 2015) to reflect the current research trends, but it would take a long time for manual cataloging to capture these changes. Furthermore, due to the lack of an objective method to measure the relatedness between journals, human cataloging is carried out based on subjective criteria, which may vary considerably from one person to another.

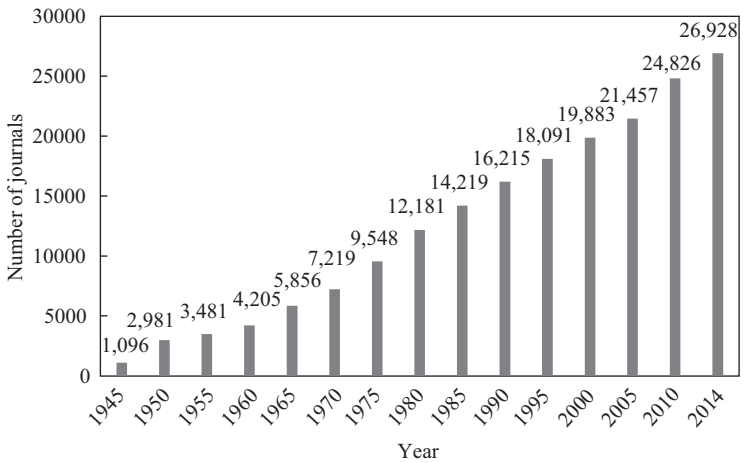

Figure 1. Number of journals in PubMed from 1945–2014.

Quantitative approaches to journal clustering and cataloging without relying on human interference have been proposed (Chen, 2008; D'Souza & Smalheiser, 2014; Eisenberg & Wells, 2014; Pudovkin & Garfield, 2002). Most studies on journal clustering use journal citation information available in Thomson Reuters' Journal Citation Reports (JCR). The citation information provides an understanding of the interaction among various scientific disciplines. Therefore two journals are likely to be related if articles published in these two journals often cite each another. For example, in the study carried out by Pudovkin and Garfield (2002) the related journal list was produced using the "relatedness factor (RF)" based on citation data in JCR. RF was calculated with the citation scores for journals that give to or receive from one journal (in their paper is *Genetics*, a core journal in the field of genetics and heredity) the highest number of citations. D'Souza and Smalheiser (2014) used three metrics to measure journal similarity based on 1) MeSH term similarity, 2) author-individuals in common between each pair of journals, and 3) articles in each journal pair written by the same author-individuals. Fujii (2007) applied link analysis techniques to the citation structures of the patent collection. After combining the citation-based scores with the text-based scores for patents, better performance than only using the text information was achieved. Although the results of the citation-based journal clustering approach are consistent with the ISI classification scheme, the full citation information is not easy to be obtained. For example, the 2013 edition of the JCR contained statistical information for approximately 8,400 science and technology journals while the number of journals in PubMed was nearly 27,000 in 2014, which means that over 2/3 journals have no citation information in JCR and hence could not be clustered using the citation-based techniques.

Another journal clustering approach based on article usage information has been proposed (Lu, Xie, & Wilbur, 2009). The idea is based on the hypothesis that if articles in two journals are often read by the same set of users, the two journals are likely to be related. It has been confirmed that the PubMed query log data (including users' searches and clicks) can be used as approximate measures of article usage so as to identify related journals. However, since the query log data is not publicly available, the usage-based journal clustering approach cannot be incorporated into the third-party applications. Moreover, both citation-based and usage-based approaches partially depend on human decisions, which introduce subjective effects into the clustering results. For example, authors tend to cite articles published in journals that they are familiar with, and article searchers may also tend to click articles of prominent journals. Therefore, journals with a lower impact factor might not be identified as related journals, even if they belong to the same research topic.

In this paper, we present a data-driven approach to mine related biomedical journals for automatic cataloging in a timely fashion. It uses the content similarity

of articles in two journals to judge the journal relatedness. This judgment is intuitive as two journals are likely to be related if they often publish papers on similar topics. The similarity between articles can be measured based on their content using term weights, e.g. using vector similarity scoring approaches (Salton & Buckley, 1988). Therefore, the clustering results of this approach are only decided by the content similarity of the journals, which are robust for automatic journal cataloging.

2 Methodology

2.1 Data Collection

The articles published from August 1, 2011 to July 31, 2012 were obtained from PubMed. These 917,844 articles belong to 4,841 unique journals. From this set of data, we first picked 740,870 articles belonging to 3,265 journals which published more than 50 papers in this time period. We then retrieved the related articles of these articles through E-utilities, which are a set of programming tools that provide a stable interface into the Entrez query and database system at the National Center for Biotechnology Information (NCBI) (NCBI, 2010). For each of the articles, we kept up to five related articles, which were displayed alongside the target articles in PubMed.

We also collected one month's PubMed query logs (from March 1, 2008 to March 31, 2008), which included a total of 8 million user sessions (after removing robot sessions) and 51 million citation retrievals. A citation retrieval is a specific MEDLINE record being clicked to display its corresponding bibliographic information and abstract text. We looked for related journals for the 3,265 journals in the 8 million user sessions. For each journal, we kept a list of the top 20 related journals.

2.2 Related Journals Identified by Article Content Similarity

How to measure the relevance of journal articles is the key to accurate retrieval of related articles. Typically most systems use TF-IDF-like schemes to determine how the articles are related. In PubMed, the relevance judgments are generated based on article content similarity, and Medical Subject Headings (MeSH) terms in MEDLINE are used for parameter estimation in the retrieval model. MeSH is a controlled vocabulary primarily used to index articles in PubMed for improving literature retrieval, and has also been used in many other scientific investigation areas (Mao & Lu, in press).

The retrieval model that underlies the related article search feature in PubMed is a topic-based content similarity model called *pmra* (Lin & Wilbur, 2007). The *pmra* model calculates the similarity between two documents using the words they have in common, with document length adjustment. It uses words from titles and abstract

as well as MeSH terms assigned to the document. The probability that the two documents are relevant given their content is estimated based on approximating the Bayesian weights for words in common. The *pmra* model uses Poisson distributions to model term frequencies within documents. The local weight of a term *t* is computed with Equation (1):

$$TF_t = (1 + e^{\alpha \times dl} \lambda^{lc-1})^{-1}, \quad (1)$$

where *dl* is the document length expressed in words, *lc* is the local frequency of *t* in document *c*, and α and λ are constants tuned to the data.

Putting the local frequency of the terms in the documents (*TF*), and the inverse document frequency (*IDF*) together, we calculate the weights of terms using Equation (2) and the document ranking function with Equation (3):

$$w_{t,c} = TF_t \times \sqrt{IDF_t} = (1 + e^{\alpha \times dl} \lambda^{lc-1})^{-1} \sqrt{IDF_t}\ , \quad (2)$$

$$Sim(c,d) = \sum_{t=1}^{N} w_{t,c} \times w_{t,d}\ . \quad (3)$$

The *pmra* model has been found to give good performance on MEDLINE documents and used to calculate "related articles" in PubMed. Based on the retrieved related articles for each article in our dataset, we compute the similarities between the journal of the original article and each of the journals of the top five related articles using the probabilities of journals' appearance in the related articles list. That is, if a journal has high probability of being displayed in the top related articles of articles in another journal, the two journals are likely related. Additionally we compute a new relevance score which uses the frequency of articles with Equation (4):

$$Rev(c,d) = \sum_{i=1}^{M} N_i(c,d)\ , \quad (4)$$

where *M* is the total number of articles published in journal *c*, $N_i(c, d)$ denotes the number of articles published in journal *d* in the related articles list of the i^{th} article of journal *c*. The use of frequency to simplify the probability estimation is reasonable although more sophisticated algorithms can be considered (e.g. taking account of the total number of papers published in the journal). To directly compare the results of the other methods, we kept the top 20 relevant journals for each journal.

2.3 Related Journals Identified through Log Analysis

The calculation of related journals is based on the existence of a set of user sessions $\{s_i\}_{i=1}^{N}$, where each user session S_i consists of a set $\{d_j^{\,i}\}_{j=1}^{n_i}$ of citation

retrievals in the form of MEDLINE records that were examined by the user during that session (Lu, Xie, & Wilbur, 2009). Let A represent a journal and t_A (s_i) denote the number of clicks through events that represent articles from journal A:

$$T_A = \sum_{i=1}^{N} t_A(s_i). \tag{5}$$

The similarity between journal A and journal B can be measured as the probability of transitioning from articles in journal A to articles in journal B: $P(B|A)$. It can be estimated as the probability of the union of three independent events: a user is looking for an article from journal A in session S_i (E_1), the article is not the last click through in the session (E_2), and the next article the user looks for in the session is from journal B (E_3). The three probabilities can be calculated with Equations (6)–(8):

$$E_1 = \frac{t_A(s_i)}{T_A}, \tag{6}$$

$$E_2 = \frac{n_i - 1}{n_i}, \tag{7}$$

$$E_3 = \frac{t_B(s_i)}{n_i - 1}. \tag{8}$$

The similarity between journal A and journal B can be computed using Equation (9):

$$P(B|A) = \sum_{i=1}^{N} \left(\frac{t_A(s_i)}{T_A} \right) \left(\frac{t_B(s_i)}{n_i - 1} \right) \left(\frac{n_i - 1}{n_i} \right) = \sum_{i=1}^{N} \left(\frac{t_A(s_i) t_A(s_i)}{T_A n_i} \right). \tag{9}$$

2.4 Evaluation Metrics

To measure the quality of retrieved related articles through the content-similarity-based approach, we use the following metrics.

2.4.1 Pearson Correlation Coefficients

To measure how well the related journals generated based on content similarity correlate with the ranking of journals, we used the article usage data in the log.

Suppose $\mathbf{X} = [x_1, x_2, \ldots, x_n]$ and $\mathbf{Y} = [y_1, y_2, \ldots, y_n]$ are a series of predicted and actual clicking numbers of n articles, respectively, the sample correlation coefficient is used to estimate the Pearson correlation coefficient r between $\mathbf{X}$ and $\mathbf{Y}$:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 (y_i - \overline{y})^2}}, \quad (10)$$

where $\overline{x}$ and $\overline{y}$ are the sample means of **X** and **Y**, and s_x and s_y are the sample standard deviations of **X** and **Y**. A value of 1 indicates a perfect positive linear relationship between **X** and **Y**.

2.4.2 Kendall's Tau Correlation Coefficients

To measure how well the content-similarity-based approach ranks the journals in the order of relevance as compared to human assessors, Kendall's tau correlation coefficients (KTCC) is used and it is defined in Equation (11):

$$\tau = \frac{n_c - n_d}{\sqrt{(n_0 - n_1)(n_0 - n_2)}}, \quad n_0 = n(n-1)/2, n_1 = \sum_i t_i(t_i - 1)/2, n_1 = \sum_j u_j(u_j - 1)/2 \quad (11)$$

where n is the number of items, n_c is the number of concordant pairs, n_d is the number of discordant pairs, t_i is the number of tied values in the i^{th} group of ties for the first ranking, u_j is the number of tied values in the j^{th} group of ties for the second ranking. A KTCC value of 1 means the approach ranked the journals in exactly the same order as human assessors and –1 means that the approach ranked the journals in exactly the opposite order of human assessors, and 0 means there is no relationship between the two orderings of the data.

2.4.3 IR Features

To measure how accurate an approach is for related journal search, we treat the approach as an information retrieval task. We specifically evaluate how the retrieved related journals have satisfied the actual goal of a user's search in terms of relevance accuracy and ranking accuracy. Therefore, the top 20 related journal search results based on usage of the original article are regarded as the golden standard, and the related journals of the original article obtained by the content-similarity-based approach are the results of the retrieval task. We measure the search accuracy using the following metrics.

(i) Precision, recall and *F*-measure: Precision is the fraction of retrieved documents that are relevant while recall is the fraction of relevant documents that are retrieved.

$$\begin{aligned}\text{Precision} &= \frac{\#(\text{relevant items retrieved})}{\#(\text{retrieved items})} = P(\text{relevant}|\text{retrieved}),\\ \text{Recall} &= \frac{\#(\text{relevant items retrieved})}{\#(\text{relevant items})} = P(\text{retrieved}|\text{relevant}).\end{aligned} \tag{12}$$

Recall and precision usually contradict each other: low precision means many results are not relevant and low recall means many relevant results are not retrieved. Therefore, they are usually combined into a single measure, such as the *F*-measure (*F*), which is the weighted harmonic mean of precision and recall and is calculated with Equation (13):

$$F = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{13}$$

This is also known as the *F*1 measures, because recall and precision are evenly weighted. In our setting, the number of retrieved results and the number of relevant journals are both 20, and thus the values of precision, recall and *F*1 are the same, which can be calculated as relevant journals retrieved and divided by 20.

(ii) Normalized discounted cumulative gain (NDCG): Given the search results of a journal, we measure how these results are accurately ranked based on the multi-level judgement of relevance (according to usage ranking) to the actual goal of a user's search task (mined from the user's behavior during the search process). NDCG is used to measure ranking accuracy (Järvelin & Kekäläinen, 2002). For every rank position *k* in the ranked list, NDCG is defined as follows:

$$NDCG(k) = \frac{1}{Z_k} \sum_{p=1}^{k} \frac{2^{S(p)} - 1}{\log(1+p)}, \tag{14}$$

where $S(p)$ is the relevance score of the document at position p in the ranked list and Z_k is a normalization factor. We assume the top 20 related journals with the original articles representing "the correct results", and each journal is judged on a scale of 1–20, with 20 for the journal which is ranked first, 1 for the journal ranked 20th.

3 Results

Journals in PubMed are assigned broad subject terms by the National Library of Medicine (NLM) to describe the journal's overall scope (Weis, 2013). All of these broad subject terms (about 120) are valid MeSH terms. We can utilize them to validate the results of the content-similarity-based approach, as journals are manually

classified into broad subjects. Here we take *Journal of the American Medical Informatics Association* (*JAMIA*) as an example.

Table 1 shows the top 10 journals related to *JAMIA* identified by the content-similarity-based approach. The top five related journals are assigned the term "Medical Informatics." Therefore, if *JAMIA* is a new journal to the system, it will be automatically cataloged into "Medical Informatics," which is exactly the broad subject term assigned to it by human indexers. While the last three journals in Table 1 are less related to this topic, the other terms in Table 1, such as "Computational Biology," "Health Services Research" and "Technology" are closely related to "Medical Informatics." This indicates that the content-similarity-based approach can cluster related journals and rank them according to their relevance to the topic.

Table 1. Top 10 journals related to *JAMIA* identified by the content-similarity-based approach.

Journal	Broad subject term(s)
*AMIA Annual Symposium Proceedings**	Medical Informatics
BMC Medical Informatics and Decision Making	Medical Informatics
Journal of Biomedical Informatics	Medical Informatics
Studies in Health Technology and Informatics	Health Services Research Medical Informatics Technology
International Journal of Medical Informatics	Medical Informatics
BMC Bioinformatics	Computational Biology
Journal of Medical Internet Research	Medical Informatics
Health Technology Assessment	Health Services Research Technology
Journal of General Internal Medicine	Internal Medicine
Pediatrics	Pediatrics
Journal of the American Medical Informatics Association (JAMIA)	Medicine

Note. * AMIA: The American Medical Informatics Association. *AMIA Annual Symposium Proceedings* is an ejournal published by AMIA annually.

We then analyze the results of the content-similarity-based approach with respect to the results of usage-based approach. We re-run the algorithm of the usage-based approach developed earlier (Lu, Xie, & Wilbur, 2009). Table 2 shows the average scores of CC, KTCC, *F*1 and NDCG with two different truncated positions 5 (NDCG@5) and 20 (NDCG@20) for journals with different number of published papers. The performance of the content-similarity-based approach is dependent on the number of papers published in a journal. For journals that published less than 300 papers in one year, those which published more papers received better performance. This is reasonable because it will be less accurate to find more related journals based on few related articles.

However, worse performance was found in journals that published more than 500 papers in one year. This is probably because the scope of those journals is broad

and they publish many papers in very diverse fields, which makes it difficult to identify related journals for them. For example, the content of *PLOS One* articles covers over 10 subject areas, from "Biology and Life Science" to "Social Science." The average *F*1 scores around 0.5 indicate that about half related journals identified by the content-similarity-based approach are really related journals according to user's search records. However, the low CC and KTCC scores (0.2617 and 0.0647 for all journals, respectively) indicate that there is much difference between the rankings of the related journals generated by the content-similarity-based approach and the rankings of the usage-based approach. On the one hand, the relatively higher NDCG@5 scores mean the top five related articles that display alongside each original article are probably published in related journals. On the other hand, the relatively lower NDCG@20 scores mean that while the "see all…" results of "similar articles" in PubMed are viewed, the full list of the related articles in the first page is less consistent with users' real information needs.

Table 2. Related journal results evaluation of the content-similarity-based approach using different metrics.

Number of papers published in 12 months	Number of Journals	CC*	KTCC*	*F*1	NDCG@5*	NDCG@20*
>50	3,265	0.2617	0.0647	0.52	0.7235	0.6654
>100	2,161	0.3271	0.1185	0.55	0.7583	0.7015
>200	1,063	0.4153	0.1767	0.59	0.7931	0.7403
>300	599	0.4646	0.1957	0.60	0.8003	0.7508
>500	233	0.4591	0.1458	0.58	0.7570	0.7161

Note. * CC: Pearson correlation coefficients; KTCC: Kendall's tau correlation coefficients; NDCG: Normalized discounted cumulative gain. NDCG@5: NDCG with truncated position 5; NDCG@20: NDCG with truncated position 20.

Figure 2 shows the distribution of the NDCG@20 values on all journals. The large number of journals that received relatively high accuracy of ranking ($0.6 < NDCG@20 < 0.9$) by the content-similarity-based approach means this approach is appropriate for most journals. The content-similarity-based approach and the usage-based method can complement each other because their results are almost identical ($NDCG@20 > 0.9$) only for a small number of journals. Furthermore, the usage of some journals might not reflect the content of the journals ($NDCG@20 < 0.6$) because users could be attracted to click these journals for some other reasons, such as high impact factor of the journal, notable or familiar authors, interesting titles, and so on.

4 Discussion

We also compared qualitatively the result lists obtained through the content-similarity-based approach and the usage-based method with the related journal list

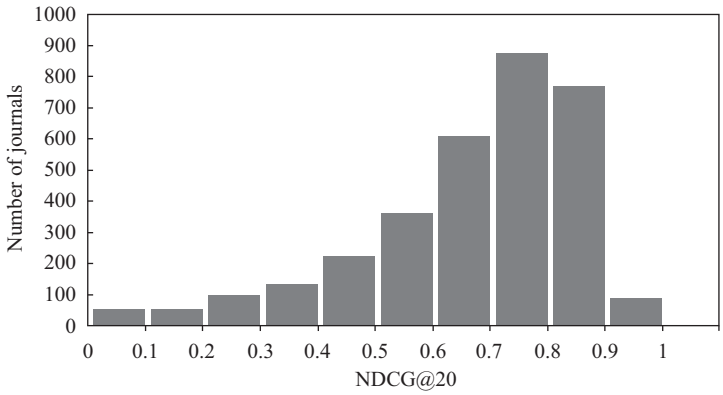

Figure 2. Distribution of the NDCG@20 values over all journals.

presented in the article of Pudovkin and Garfield (2002), where the list was produced using citation data. Given the journal *Genetics*, Table 3 shows the top 20 related journals identified by the citation-based, usage-based and content-similarity-based approach, respectively. In general, the results of the three methods overlap each other in large part. Similar to *Genetics*, most of the journals publish articles in the field of genetics and heredity. Although some journals cover more full range of scientific disciplines, such as *Science*, *Nature* and *PNAS*, they mainly focus on the life science field rather than other fields such as computer science.

We further examined the most frequent MeSH terms assigned to the articles published in these 32 journals (including *Genetics*). The Check tags (a special set of MeSH headings that are mentioned in almost every article such as human, animal, male, female, child, etc.[①] have the highest frequency over all journals in PubMed, but are not useful for journal cataloging. The top three MeSH terms (after excluding the Check tags) in articles of each journal are shown in Table 4, in which most MeSH terms are related to the field of genetics and heredity, including the most frequently assigned MeSH terms: gene expression, gene expression regulation, and DNA. Since MeSH terms are assigned to articles by human indexers, we can conclude that most journals identified by the three methods are correctly related to the given journal.

We also found the ranking based on citation count has high correlation with the ranking based on usage. This is reasonable because the high clicking number of an article would lead to high citation count of that article, and such phenomenon has been explored in previous work (Brody, Harnad, & Carr, 2006). It should also be

① http://www.nlm.nih.gov/bsd/indexing/training/CHK_010.html

Table 3. Top 20 journals related to *Genetics* identified by the citation-based, usage-based and content-similarity-based method, respectively.

Citation-based	Usage-based	Content-similarity-based
PNAS	*PNAS*	*PLOS Genetics*
Cell	*JBC*	*PNAS*
Nature	*Nature*	*MCB*
MCB	*Science*	*PLOS One*
Science	*MCB*	*Genome Research*
TAG	*Cell*	*Nature*
Evolution	*Development*	*Eukaryotic Cell*
EMBO Journal	*Genes & Development*	*Evolution*
Genes & Development	*NAR*	*MBoC*
NAR	*EMBO Journal*	*Molecular Ecology*
JBC	*Current Biology: CB*	*JBC*
MBE	*MBoC*	*Science*
Journal of Bacteriology	*Developmental Biology*	*Developmental Biology*
MGG	*MBE*	*Cell*
Heredity	*Nature Genetics*	*Current Biology: CB*
Development	*JCB*	*Genes & Development*
Genetical Research	*Journal of Bacteriology*	*Development*
Genome	*CCM*	*MBE*
JMC	*PLOS Genetics*	*Heredity*
JCB	*AJHG*	*AJHG*

Note. *PNAS*: *Proceedings of the National Academy of Sciences of the USA*; *JBC*: *Journal of Biological Chemistry*; *MCB*: *Molecular and Cellular Biology*; *TAG*: *Theoretical and Applied Genetics*; *NAR*: *Nucleic Acid Research*; *MBoC*: *Molecular Biology of the Cell*; *MBE*: *Molecular Biology and Evolution*; *MGG*: *Molecular & General Genetics*; *JCB*: *Journal of Cell Biology*; *CCM*: *Critical Care Medicine*; *JMC*: *Journal of Molecular Biology*; *AJHG*: *American Journal of Human Genetics*.

pointed out that the latency between the publishing and the peak citing time of an article is longer than the latency between its publishing and peak clicking time (Mao & Lu, 2013), but the time lag probably is not long enough for changing the scope of a journal. Furthermore, we found that the ranks of the journals that not only focus on the field of genetics and heredity were lower in the content-similarity-based approach than in the other two approaches. Such differences were also observed by D'Souza and Smalheiser (2014). This is probably because these journals, such as *Science* and *Nature*, have quite high impact factor, and they tend to receive more citations and clicks. Therefore, the results of the content-similarity-based approach were not subject to human factors, since some journals with lower impact factor could be more related to this field.

5 Conclusion

This research is valuable in using article content similarity to explore the proximity pattern of biomedical journals, validating that the content-similarity-based approach is useful in clustering related journals.

Table 4. Top 3 MeSH terms in the articles of the journals identified by the three methods.

Journal	MeSH #1	MeSH #2	MeSH #3
Genetics	Mutation	Models, genetic	Genome
PNAS	DNA	Gene expression	Molecular sequence data
Cell	RNA	DNA	Molecular sequence data
Nature	Research	Research personnel	DNA
MCB	Gene expression	Cell line	Gene expression regulation
Science	DNA	Science	RNA
TAG	Chromosome mapping	Quantitative trait loci	Genes
Evolution	Evolution, molecular	Biological evolution	Selection, genetic
EMBO Journal	DNA	RNA	Gene expression
Genes & Development	Gene expression	Gene expression regulation	Cell line
NAR	DNA	RNA	Internet
JBC	Proteins	Gene expression regulation	Protein binding
MBE	Evolution, molecular	Phylogeny	Genome
Journal of Bacteriology	Bacteria	Gene expression	Gene expression regulation
MGG	Gene expression	Gene expression regulation	DNA
Heredity	Genetic variation	Genetics	Genetics, population
Development	Gene expression	Gene expression regulation	Gene expression regulation, developmental
Genetical Research	Models, genetic	Genotype	Chromosomes mapping
Genome	Phylogeny	Genes	Chromosomes
JMB	Models, molecular	Protein binding	Protein confirmation
JCB	Cell line	Protein transport	Cells
Current Biology: CB	Drosophila	Biological evolution	Gene expression
MBoC	Protein transport	RNA	Protein binding
Developmental Biology	Gene expression	Gene expression regulation	Gene expression regulation, developmental
Nature Genetics	Genome	Mutation	Polymorphism, single nucleotide
CCM	Intensive care	Intensive care units	Critical illness
PLOS Genetics	Gene expression	Gene expression regulation	DNA
AJHG	Mutation	Genetic predisposition to disease	Pedigree
PLOS One	Gene expression	Gene expression regulation	Cell line
Genome Research	Genome	Gene expression	DNA
Eukaryotic Cell	Fungal proteins	Gene expression	Gene expression regulation
Molecular Ecology	Genetic variation	Genetics, population	Genetics

Note. *PNAS*: *Proceedings of the National Academy of Sciences of the USA*; *MCB*: *Molecular and Cellular Biology*; *TAG*: *Theoretical and Applied Genetics*; *NAR*: *Nucleic Acid Research*; *JBC*: *Journal of Biological Chemistry*; *MBE*: *Molecular Biology and Evolution*; *MGG*: *Molecular & General Genetics*; *JMB*: *Journal of Molecular Biology*; *JCB*: *Journal of Cell Biology*; *MBoC*: *Molecular Biology of the Cell*; *CCM*: *Critical Care Medicine*; *AJHG*: *American Journal of Human Genetics*.

Further analysis also produces several insights. First, the clustering results based on content similarity, Web usage and citation have high correlation with each other, and are consistent with the results of manual cataloging. Second, results of the content-similarity-based approach are considerably less subject to human factors and some journals with lower impact factors can be clustered and ranked higher in the related journal list.

In conclusion, this research offers another way of clustering biomedical journals by using article content similarity information, other than widely used journal citation information. Moreover, incorporating the usage and citation information of journals will probably improve the accuracy of journal clustering. We would like to investigate this issue in the future and extend this work with other related investigation research, such as Klavans and Boyack's work (2006).

Acknowledgements

We would like to thank Dr. John Wilbur for his helpful discussion on this project. This research is supported by NIH Intramural Research Program, National Library of Medicine.

Author Contributions

Y.Q. Mao (mao.yuqing@msn.com) implemented the methods and performed the experiments, analyzed the results, and wrote the first draft. Z.L. Lu (luzy@ncbi.nlm.nih.gov) conceived the project. Both authors participated in the design, results discussion and writing of the paper. Both authors read and approved the final manuscript.

References

Brody, T., Harnad, S., & Carr, L. (2006). Earlier web usage statistics as predictors of later citation impact. Journal of the American Society for Information Science and Technology, 57(8), 1060–1072.

Chen, C. M. (2008). Classification of scientific networks using aggregated journal-journal citation relations in the Journal Citation Reports. Journal of the American Society for Information Science and Technology, 59(14), 2296–2304.

D'Souza, J. L., & Smalheiser, N. R. (2014). Three journal similarity metrics and their application to biomedical journals. PLOS One, 9(12), e115681.

Eisenberg, T., & Wells, M. T. (2014). Ranking law journals and the limits of journal citation reports. Economic Inquiry, 52(4), 1301–1314.

Fujii, A. (2007). Enhancing patent retrieval by citation analysis. In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. (pp. 793–794). New York: ACM.

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems (TOIS), 20(4), 422–446.

Kang, N., Doornenbal, M. A., & Schijvenaars, R. J. (2015). Elsevier journal finder: Recommending journals for your paper. In Proceedings of the 9th ACM Conference on Recommender Systems. (pp. 261–264). New York: ACM.

Klavans, R., & Boyack, K. W. (2006). Identifying a better measure of relatedness for mapping science. Journal of the American Society for Information Science and Technology, 57(2), 251–263.

Lin, J., & Wilbur, W. J. (2007). PubMed related articles: A probabilistic topic-based model for content similarity. BMC Bioinformatics, 8(1), 423.

Lu, Z., Xie, N., & Wilbur, W. J. (2009). Identifying related journals through log analysis. Bioinformatics, 25(22), 3038–3039.

Mao, Y., & Lu, Z. (2013). Predicting clicks of PubMed articles. In AMIA Annual Symposium Proceedings. (pp. 947–956). Bethesda, Maryland: American Medical Informatics Association.

Mao, Y., & Lu, Z. (in press). MeSH now: Automatic MeSH indexing at PubMed scale via learning to rank. Journal of Biomedical Semantics.

Pudovkin, A. I., & Garfield, E. (2002). Algorithmic procedure for finding semantically related journals. Journal of the American Society for Information Science and Technology, 53(13), 1113–1119.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information Processing & Management, 24(5), 513–523.

Shultz, M. (2007). Comparing test searches in PubMed and Google Scholar. Journal of the Medical Library Association: JMLA, 95(4), 442.

Small, H. G., & Koenig, M. E. (1977). Journal clustering using a bibliographic coupling method. Information Processing & Management, 13(5), 277–288.

Srivastava, C.V., Towery, N.D., & Zuckerman, B. (2007). Challenges and opportunities for research portfolio analysis, management, and evaluation. Research Evaluation, 16(3), 152–156.

Weis, S. (2013). NLM Catalog. Retrieved from http://www.ncbi.nlm.nih.gov/books/NBK153380/.

The National Center for Biotechnology Information (NCBI). (2010). Entrez programming utilities help. Retrieved from http://www.ncbi.nlm.nih.gov/books/NBK25501/.