Firework Visualization: A Model for Local Citation Analysis

Si-Chi Chin
Information Science
The University of Iowa
Iowa City, Iowa 52242
si-chi-chin@uiowa.edu

Charisse Madlock-Brown
Health Informatics
The University of Iowa
Iowa City, Iowa 52242
charisse-madlock-brown@uiowa.edu

W. Nick Street
Management Sciences Dept.
The University of Iowa
Iowa City, Iowa 52242
nick-street@uiowa.edu

David Eichmann
Institute for Clinical and Translational Science (ICTS)
The University of Iowa
Iowa City, Iowa 52242
david-eichmann@uiowa.edu

ABSTRACT

Citation chasing, the pursuit of references from one publication to another, is a popular technique among researchers for retrieving relevant and related literature. While a focused and precise method, citation chasing may be incomplete and unstable. Different choices of paths can lead to different sets of literature and generate inconsistent search results. Revealing the linkages among references would guide citation chasing as well as improve and stabilize the results since it enables researchers to develop search strategies.

In this paper we use network analysis to examine the connections among papers cited as references, identifying and investigating papers with high authority and hub values. We construct local citation networks, introducing the firework visualization model to support citation chasing. We further introduce an interactive browser, designed to provide direct manipulation of references and aggregated summaries of retrieved results. The proposed firework model allows users to intuitively browse a local citation network and find relevant documents efficiently.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Search process; H.3.4 [**Systems and Software**]: Information networks; J.3 [**Life and Medical Sciences**]: Medical information systems

General Terms

Design, Documentation, Human Factors

IHI'12, January 28–30, 2012, Miami, Florida, USA.

Keywords

Firework model, citation chasing, information interaction, local citation network

1. INTRODUCTION

Finding the right articles has become increasingly more difficult as the number of articles published each year increases rapidly. The number of articles that were indexed per year by NIH's MEDLINE citation database increased from 361,000 in 1990 to 741,000 in 2010. Because of systems like PubMed and Google Scholar, researchers have consistently been able to engage in "strategic reading," a technique using online databases to "search, filter, scan, link, annotate, analyze fragments of content" [17], to scan more articles and pinpoint relevant sections. Strategic reading usually extends beyond query-based search. Researchers exploit links among papers provided by systems to retrieve related and relevant information, whether through keyword tags or related papers lists. Citation chasing, as another common strategic reading method, is the process of using both reference lists and lists of citing articles for a particular paper to identify new information. Citation chasing is useful because simple manual search may fail to return relevant information as users struggle to construct precise search queries to represent their information need.

Linking methods, such as citation chasing, are common practices among researchers fulfilling their information needs. Research has suggested that linking is a preferred search strategy in the third region of Bradford's distribution, as shown in Figure 1 [2]. Bradford's law divides journals in a field into three regions, indicating that the number of journals (information sources) that must be accessed increases exponentially in order to retrieve the same number of relevant papers as one moves away from the core region. As shown in Figure 1, linking methods dominate when researchers are working with a large number of less relevant journals because they keep the search narrow and deep.

Citation chasing is an intuitive and useful linking method. Scientists often use citation chasing when doing research as it allows them to find articles relevant to those they have already selected, which assists in the obtaining of an overview

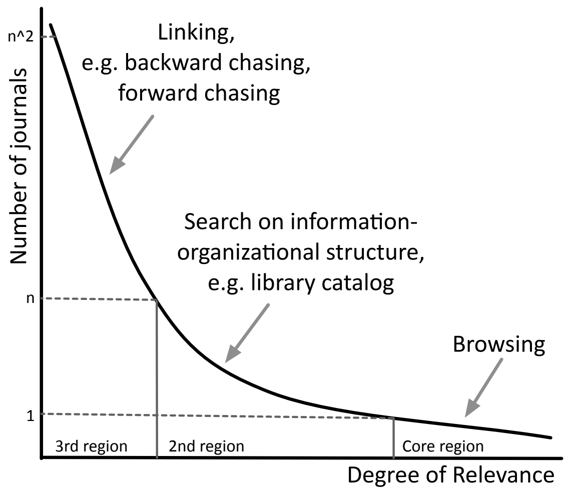


Figure 1: Optimum searching techniques for each of the three Bradford regions. Bates indicates that linking is a prime searching approach when a researcher works with a large number of less relevant journals. Inspired by [2].

of the literature in the research field [6, 9, 17, 16]. In the interrelated fields of Infometrics, Bibliometrics, and Scietomentrics, researchers also have used citation information for years to generate overviews, and to map and chart changes in scientific fields [12, 1].

However, chasing citations manually, even on a local scale, becomes a tedious task as the reference list grows. As shown in Table 1, more than 45% of the articles in the MEDLINE database contain 50 or more references. If a researcher starts with five seed articles in a subject, there are likely be more than 200 papers from the reference list upon which to initiate citation chasing. This demands a systematic approach to analyze the retrieved citations to guide linking strategies.

Table 1: Distribution of reference count in MEDLINE. More than 45% of the articles have more than 50 references.

Ref count	Freq	Perc
>= 50	695,960	45.7 %
40 - 50	159,217	10.4 %
30 - 40	184,758	12.1 %
20 - 30	207,448	13.6 %
10 - 20	199,029	13.1 %
< 10	74,922	4.9 %
Total	1,521,334	

Information visualization has gained increasing attention as it provides users an overview of search results to support informed decisions about how to pursue their searches. Several systems such as GoPubMed [1] [4], Texpresso [15], and iHOP [11] use predefined categories or classification techniques to group resources or create links between papers. The main limitation of these systems is that the links and categories are only as good as the ontology used. The most important categories for a given set of search results may not always appear in an ontology, because they may be new or too specific. Another attempt from Lin et al. [14] uses clustering to generate categories dynamically. The authors use clustering and ranking to provide users with meaningful categories based on keywords and sentences in the result set. Their system generates a relevance list within the clusters to further assist users in identifying relevant papers. While these systems are useful, none of them make use of the linkage between citations to allow users to find new results without having to construct a specific query. We believe visualizing local citation networks will improve a user's ability to find relevant information.

Local citation analysis, as we define it, explores the interconnection among references listed in a paper and provides summaries of a selected subset of the cited papers. We use the term "local citations" to refer to papers cited in the reference list. In contrast to global citation analysis, which counts the "cited by" links of a paper using the entire available document space, local citation analysis narrows the scope to one paper – the paper of interest. Local citation analysis aims to help researchers quickly grasp the scope and the landscape of a paper at a higher level.

Network visualizations have been used for decades to represent relationships among data. However, as the graph becomes dense users have a difficult time distinguishing between nodes and edges and comprehending the network [10]. When links overlap it becomes difficult to count links in an area of the graph, and observe link length, thus limiting to usefulness of the visualization. Several researchers attempt to circumvent these problems using sophisticated layout managers that dynamically determine how nodes will be placed based on the complexity of the graph [5]. For instance, a 3D representation can help avoid link occlusion, but not completely [8]. Matrix-based representations of networks have also been used, for which there is no problem of occlusion. Though matrix-based representations may be useful, comparisons of the readability of matrix-based representations to node-link representations demonstrate that for small graphs, the node-link model is more readable [8]. We utilize local citation analysis as opposed to global citation analysis to limit the representation to a more manageable set of nodes and links to improve user understanding of the network.

In this paper, we construct and analyze local citation networks, investigating the long-tail distribution property of citations and identifying papers with high authority and hub values (see Section 2). We further propose a firework visualization model and an interactive browser, designed to provide direct manipulation of references and aggregated summaries of retrieved results (see Section 4). The goal of the system is to support information interaction and hence to facilitate search. To our knowledge, it is the first work attempting network visualization in support of footnote chasing.

2. LONG-TAIL DISTRIBUTION

In this section we investigate the long tail property, a statistical property describing a distribution skewed to the right, for two survey papers published in 2000 on the occurrence count for authors, journals, and Medical Subject

[1] http://www.gopubmed.org/

Headings (MeSH) [2]. The long tail of the frequency distributions informs researchers who are the distinguished scholars of the field, what are the most important journals, and what are the dominant MeSH terms. By examining the long tail, a novice researcher would be able to grasp the broad but prominent outline of a given field. An expert researcher may compare the landscapes of different but related paper by their frequency distributions. Moreover, the less frequent MeSH terms may indicate future research directions.

Table 2 summarizes the two survey articles retrieved from 2011 MEDLINE/PubMed Baseline Distribution [3]. Both articles cited more than three hundred papers in their reference lists and have more than six hundred distinct cited authors. In contrast to Paper 2, the cited references in Paper 1 concentrate more on a few primary journals of the field and have higher inter-connectivity among the papers.

Table 2: Case study: the description of the seed article.

	Paper 1	**Paper 2**
PMID	10684922	10974125
Title	Survey and summary: transcription by RNA polymerases I and III.	Signal peptide-dependent protein transport in Bacillus subtilis: a genome-based survey of the secretome.
Pub year	2000	2000
Author count	632	623
Journal count	29	48
MeSH count	498	617
Ref count	351	338
Link count	2029	976

We observe the long tail property when we plot the frequency distribution of the number of occurrences of authors, journals, and MeSH terms. Figure 2 shows the three frequency distributions for the two example articles. All six charts demonstrate two common characteristics: first, the points at the bottom right corners reflect the few instances that occur more often in the reference list; second, the points at the top left corner reflect the majority of instances occurring only once in the references. For example, the left column shows the frequency distribution on authors. The two points at the bottom right corner reflect the two most popular authors in the reference list. For example, Grummt authored 27 papers and Sentenac authored 23 papers out of the 351 references, whereas the majority of the authors (414 out of 632) at the top left corner occur only once in the reference list. Similar patterns exist for the frequency distribution on journals and MeSH terms. For both papers, more than 200 MeSH terms occur only once. On the other hand, dominant MeSH terms appear on the bottom right. Among the 29 journals cited in the reference list, the top 2 journals (*Molecular and Cellular Biology* and *Cell* for Paper 1) comprise more than a third of the cited papers.

Figure 3 shows the in-link and out-link frequency distributions for local citations. The number of in-links of a targeted paper A is the number of other papers that cite A. The number of out-links is the number of citations existing in the targeted paper. We observed the long tail property for the local in-link distributions for both papers. However, the long-tail pattern is unclear for local out-link distributions. The end of the long tail indicates papers with higher in-degree or out-degree. Papers with only one or few in-links or out-links gather at the top-left corner.

3. AUTHORITY AND HUB

In network analysis, hubs and authorities are commonly used to assess the importance of nodes [13]. In the case of citation analysis, a node represents a paper and the the edges are directed links pointing to other papers.

Prior research has used network analysis methods to study the connections among citations. For example, Yin et al. [18] used linkage-based metrics to rank document importance to enhance retrieval performance on the TREC 2007 Genomics dataset. The authors indicated that InDegree is a suitable metric for biomedical literature retrieval. Fowler and Jeon [7] used hub and authority metrics to analyze U.S. Supreme Court opinions. They used network analysis methods to investigate the influences of Supreme Court opinions. Our work differs in that the goal of our proposed system is to support user search behaviors (e.g. citation chasing) as opposed to developing a document recommender system that performs the search for users.

The computation of hub and authority scores is iterative. A paper with high authority score is a paper widely cited by papers with higher hub scores; a hub citation is a paper having citations with higher authority scores. The hub and authority metrics are closely related to the in-degree (number of citations that cite a paper) and the out-degree (number of papers cited in a paper) of a citation. However, one should consider a paper cited by influential papers to be more important. The in-degree value neglects the varying importance of in-link citations. Similarly, the out-degree of a paper also neglects the fact that a paper citing more influential papers should be considered important. Hub and authority metrics fill in the missing perspectives in in-degree and out-degree.

Figure 4 presents the distribution of authority and hub, authority and in-degree, and hub and out-degree for the two selected papers. The numbered labels in the figure are the PMID of citations. In general, the distribution patterns of hub-authority pairs are unclear. Although a linear relationship appears to exist for authority and in-degree, and hub and out-degree, the pattern is not consistent for the two papers. From subgraph (b) and (d), we note that papers with high in-degree are not always papers with high authority scores. Meanwhile, subgraph (c) also indicates that papers with high out-degree may not have high hub scores because they contain fewer papers with high authority scores. A hub paper with higher authority value (e.g. the paper "8617241" in (a) and the paper "8096622" in (d)) might be more useful since it has higher local citations.

4. FIREWORK VISUALIZATION MODEL

We create the Firework visualization model to represent inter-connections among local citations. The model is in-

[2] MeSH is a controlled vocabulary to index literature of life sciences, such as journal articles and books. See http://www.nlm.nih.gov/mesh/ for more details.

[3] http://www.nlm.nih.gov/bsd/licensee/2011_stats/baseline_doc.html

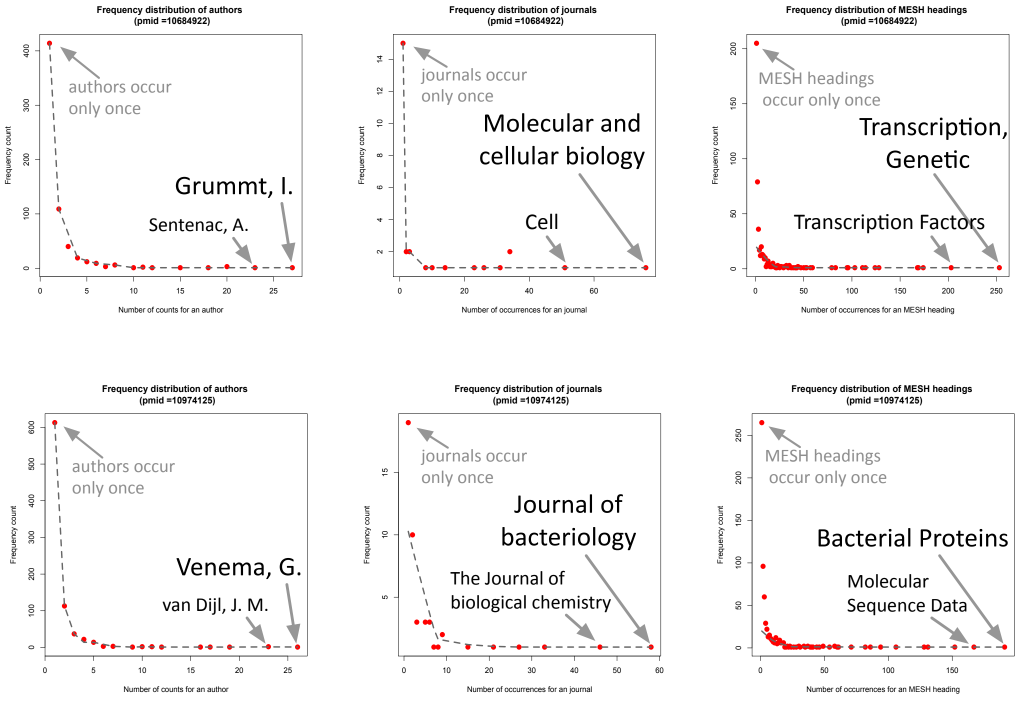


Figure 2: Frequency distribution for the number of occurrences on authors, journals, and MeSH terms. The x-axis is the number of occurrences for an author, journal, and MeSH and the y-axis is the frequency count for the number of occurrences. For example, the top-left sub-graph shows that there are more than four hundred authors occurring only once and there exists only one author, Grummt, cited 27 times in the local citation analysis. We observe the long tail characteristics for the three frequency distributions for both example articles.

spired by the process and the shape of a fireworks display. The goal of the model is to allow users to scan a paper strategically and to assist the user in developing and adjusting their citation chasing strategies. Figure 5 shows an example of a single *floral break*, a terminology commonly used in firework displays, a spherical display with a chosen paper at the center surrounded by and pointed to by papers that cited it. A firework display can have one or more *launch positions*, representing the center papers chosen by users. The number of launch positions equals the number of floral breaks. Figure 6 shows an example of a double floral break with two launch positions. Similar to constructing queries, users specify launch position(s) that determine the shape of floral break(s) to navigate the local citation space.

Figure 5 and Figure 6 are prototypes of a proposed interactive browser. The floral break was was implemented using the Java Universal Network/ Graph (JUNG) [4] Framework, applying the Fruchterman-Reingold (FR) graph layout algorithm to determine the distances between nodes. The nodes in the floral breaks are color-coded based on their in-degree. Red nodes are the top five citations with highest in-degree. Citations ranked from 6 to 10 are green and nodes with in-degree higher than 15 but not in the top 10 are color-coded in blue. The launch position in Figure 5 is the citation of highest local in-degree for Paper 1 (described in Section 2), the paper at the tip of the long tail in Figure 3. The proposed system prototype presents on the right side a panel with the summaries of the distribution of authors, publication years, journals, MeSH term frequency, and local MeSH tf-idf (see Section 4.2). When a user mouses over a node, a short dialogue window would emerge to provide a basic description of the corresponding linked paper. Assessing the floral break provides insight to users whether the chosen launch position might be the next hop in citation chasing.

Figure 6 is an example of dual floral breaks. The firework model shows a distinct cluster of the co-citations of the two launch positions. The interactive browser provides selection functionality (the oval dashed lines) and presents its aggregated summaries on the right. Another extension of the interactive browser is reflected in the consecutive views of floral breaks shown in Figure 7. The figure shows a consecutive view for the top 5 local citations with highest in-degree values. The blue and orange arrows always point at the same articles. From the multiple displays, we observe the relative positions of the two targeted papers and the new added citations. The shorter distances between the multiple launch positions (the red nodes) implies that they are similar because of the number of co-citations.

Using floral breaks as a representation will allow users to easily grasp the semantics of the relationship between cited document collections. Using this model the difference be-

[4] JUNG is an open-source software library used for the analysis, and visualization of networks, see http://jung.sourceforge.net/

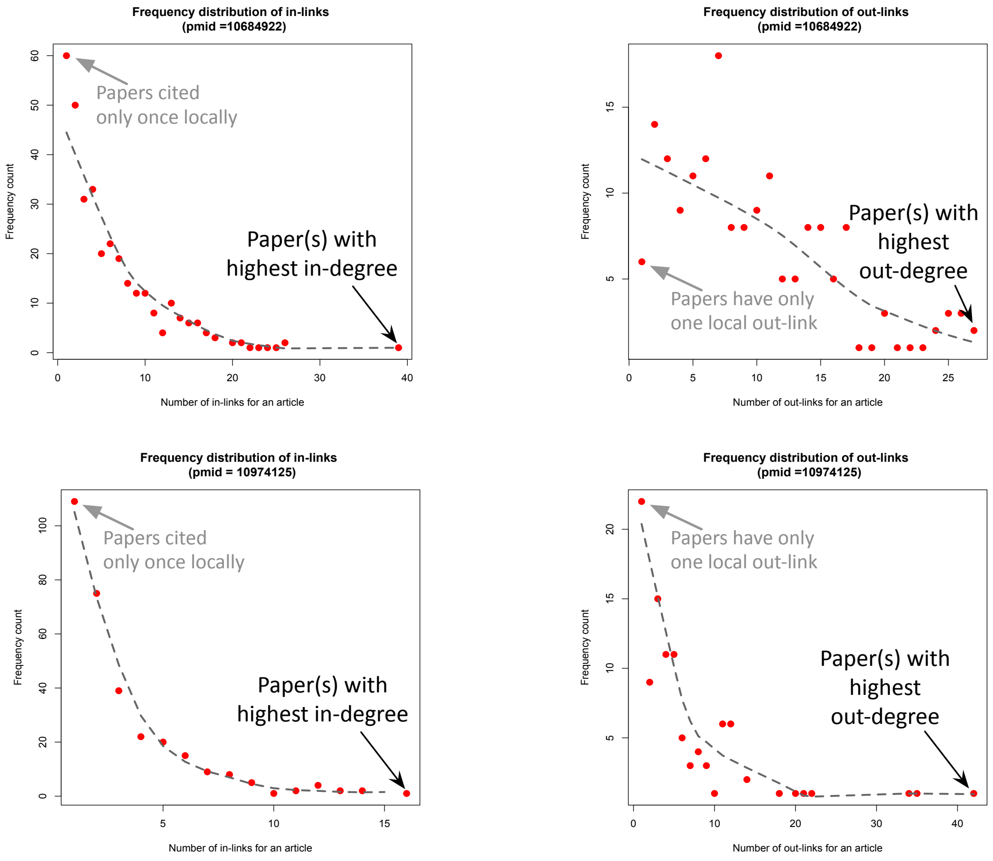


Figure 3: Frequency distribution of in-links and out-links for two articles (pmid = 10684922 and pmid = 10974125). Both graphs for in-link distributions follow a power-law, with a long tail to the right. For in-link distribution graphs, the tails contain papers that are most locally cited, that is, papers with higher authority scores. Most papers in the reference list have only one or two local citations. For out-link distribution graphs, the papers in the tail have higher hub scores. However, the power-law distribution feature is less evident for the out-link distributions.

tween seed nodes and document nodes is clear and users can easily identify overlap between cited document collections.

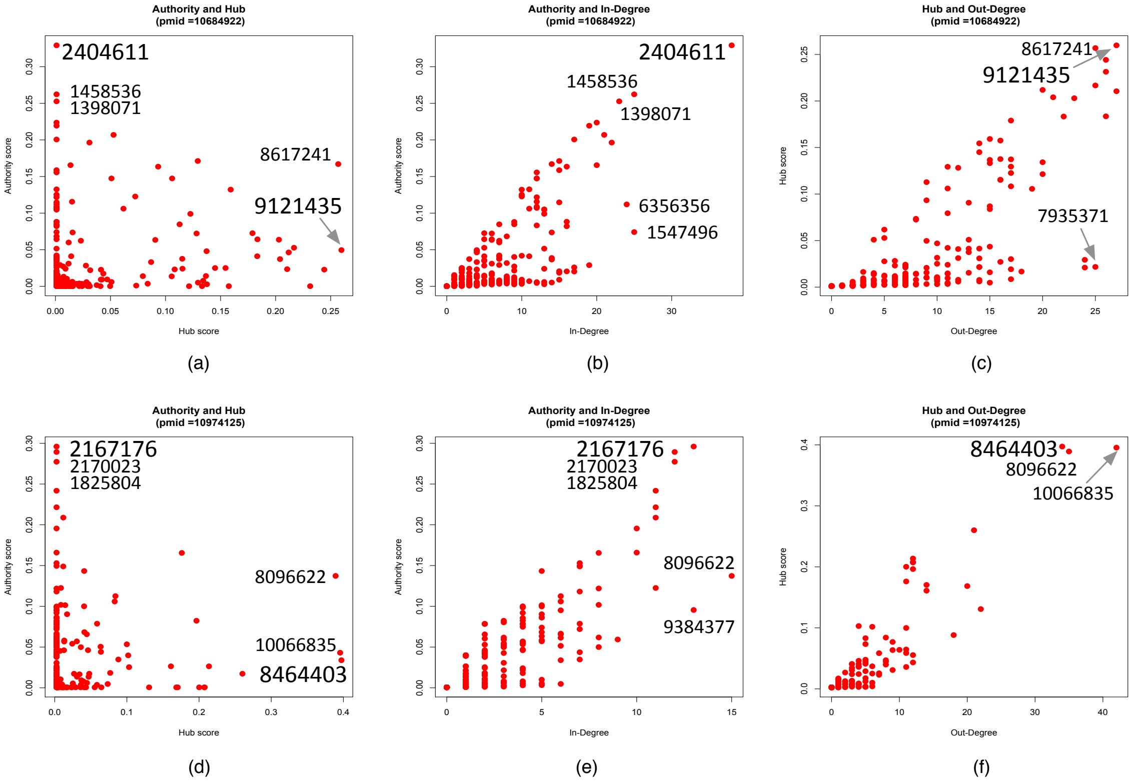
4.1 System process flow model

MEDLINE is a rich biomedical bibliographic database heavily used by health science researchers. It contains over 19 million citations from over 4,600 journals, providing a substantial resource in support of research on information retrieval, text mining, and natural language processing. The National Library of Medicine (NLM) distributes MEDLINE in eXtensible Markup Language (XML)-formatted text files. We loaded and parsed the XML files into a relational database system.

Figure 8 describes the infrastructure of our proposed firework visualization model. The circular solid arrows show the iterative features of the system. To start the iterative process, the system selects an initial launch position based on heuristics or simply at random. Once a launch position is located, a .net formatted file is extracted from the database system and then used to display one or more floral breaks. Researchers browse the network visualization, investigate the aggregated search result, and navigate the space with a control panel. The interactive browser is used to locate a new launch position – a new query. The updated query is converted into SQL and sent to the database system. The system then re-populates the network data to start another cycle of search.

4.2 Local MeSH Tf-idf

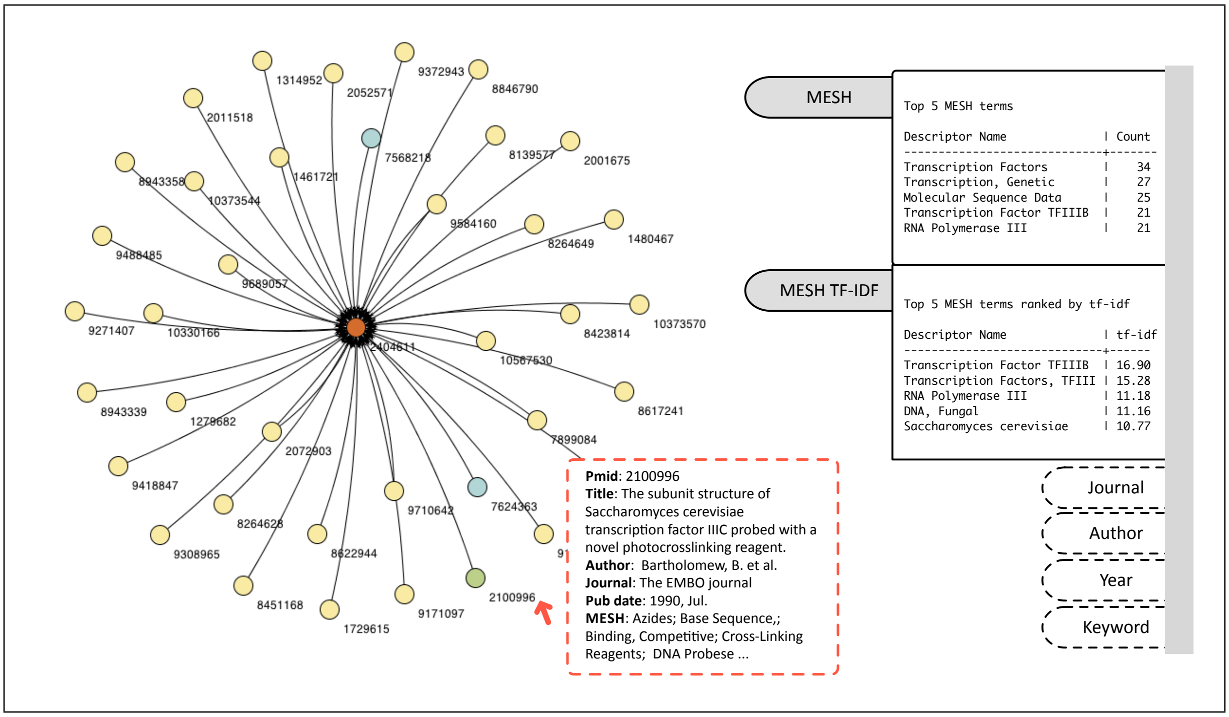
In this section, we describe how to compute MeSH tf-idf for local citation analysis. In order to distinguish the properties of displayed floral breaks from the rest of local citations, we create and compute *local MeSH tf-idf* for the retrieved subset of papers. The goal of local MeSH tf-idf is to identify the distinct MeSH terms that best describe the displayed floral break by weighting MeSH term frequencies (TF) with inverse document frequencies (IDF). We use $tf_{i,J}$ to denote the term frequency for the MeSH term i in the floral break J, where J is a set of documents, and compute



(a) (b) (c)

(d) (e) (f)

Figure 4: Authority, Hub, and Degree

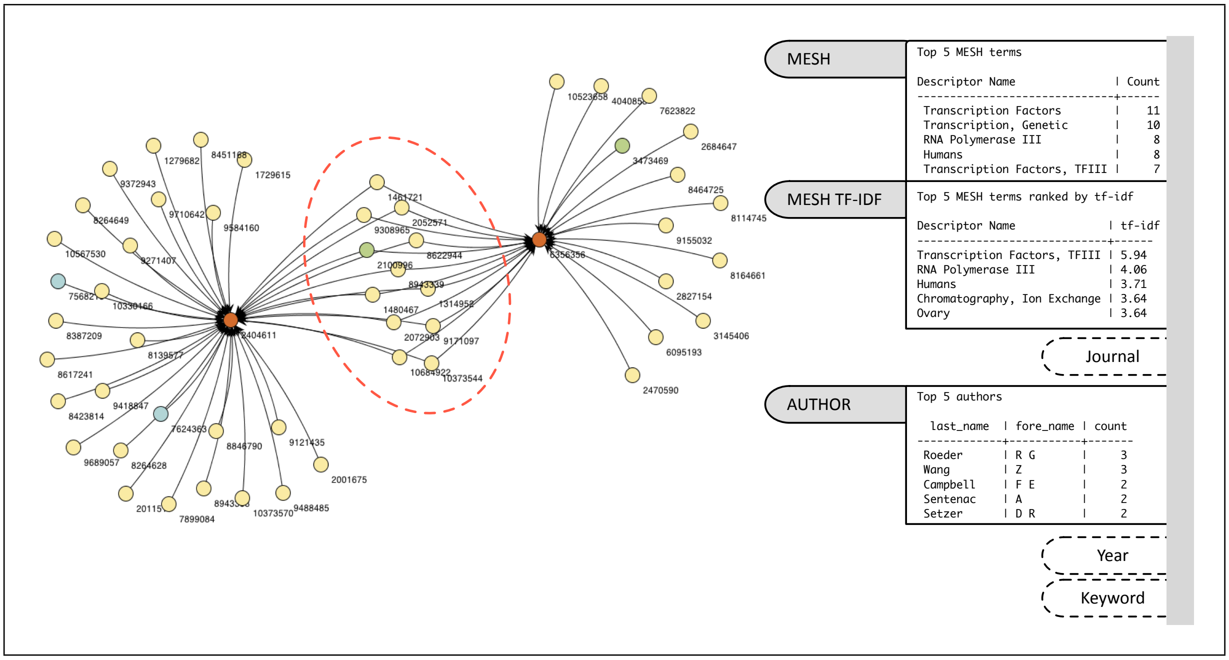


```
Top 5 MESH terms

Descriptor Name                    | Count
-----------------------------------+-------
Transcription Factors              |    34
Transcription, Genetic             |    27
Molecular Sequence Data            |    25
Transcription Factor TFIIIB        |    21
RNA Polymerase III                 |    21
```

```
Top 5 MESH terms ranked by tf-idf

Descriptor Name                     | tf-idf
------------------------------------+------
Transcription Factor TFIIIB         | 16.90
Transcription Factors, TFIII        | 15.28
RNA Polymerase III                  | 11.18
DNA, Fungal                         | 11.16
Saccharomyces cerevisiae            | 10.77
```

Pmid: 2100996
Title: The subunit structure of Saccharomyces cerevisiae transcription factor IIIC probed with a novel photocrosslinking reagent.
Author: Bartholomew, B. et al.
Journal: The EMBO journal
Pub date: 1990, Jul.
MESH: Azides; Base Sequence,; Binding, Competitive; Cross-Linking Reagents; DNA Probese ...

Figure 5: User interface prototype example I. PMID = 2404611

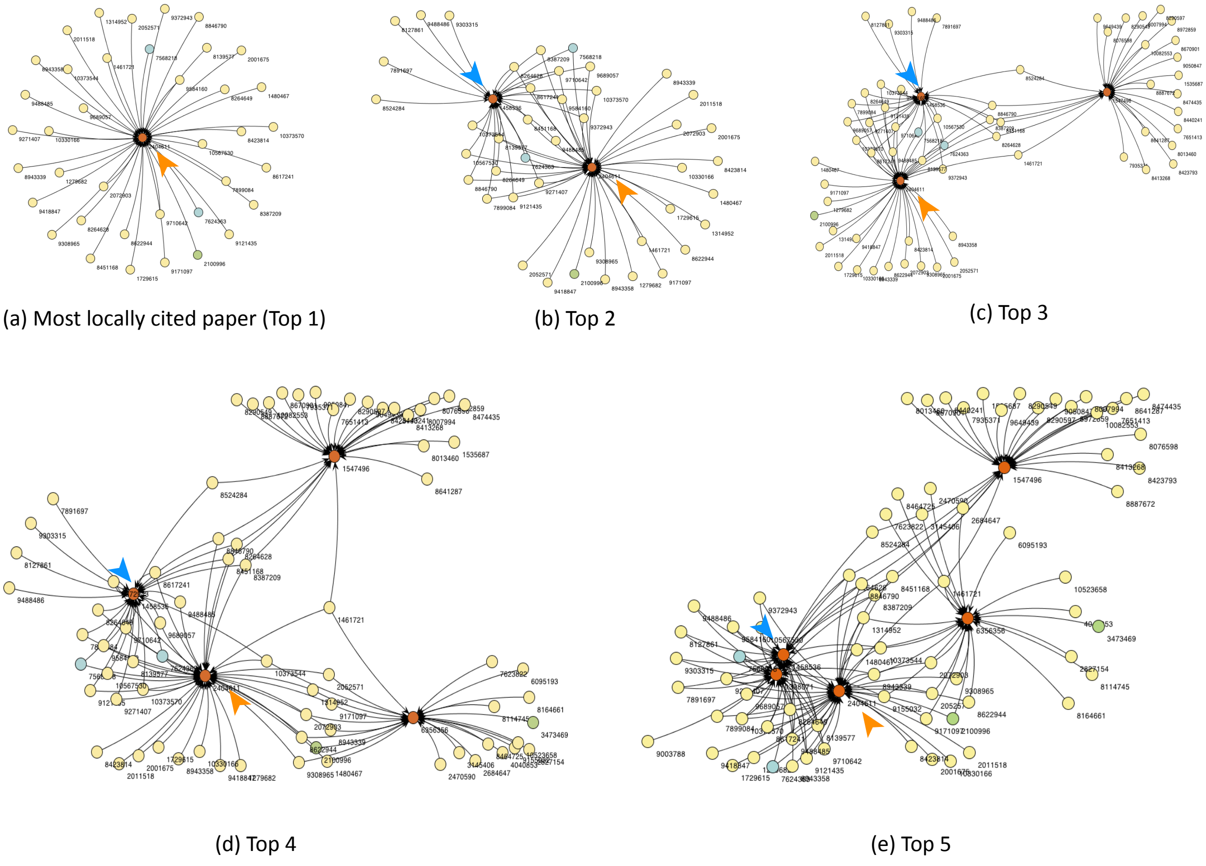


```
Top 5 MESH terms

Descriptor Name                    | Count
-----------------------------------+------
 Transcription Factors             |    11
 Transcription, Genetic            |    10
 RNA Polymerase III                |     8
 Humans                            |     8
 Transcription Factors, TFIII      |     7
```

```
Top 5 MESH terms ranked by tf-idf

Descriptor Name                    | tf-idf
-----------------------------------+------
Transcription Factors, TFIII | 5.94
RNA Polymerase III           | 4.06
Humans                       | 3.71
Chromatography, Ion Exchange | 3.64
Ovary                        | 3.64
```

```
Top 5 authors

 last_name   | fore_name  | count
-------------+------------+-------
 Roeder      | R G        |     3
 Wang        | Z          |     3
 Campbell    | F E        |     2
 Sentenac    | A          |     2
 Setzer      | D R        |     2
```

Figure 6: User interface prototype example III. PMID = 2404611, 6356356



(a) Most locally cited paper (Top 1)

(b) Top 2

(c) Top 3

(d) Top 4

(e) Top 5

Figure 7: Consecutive floral breaks.

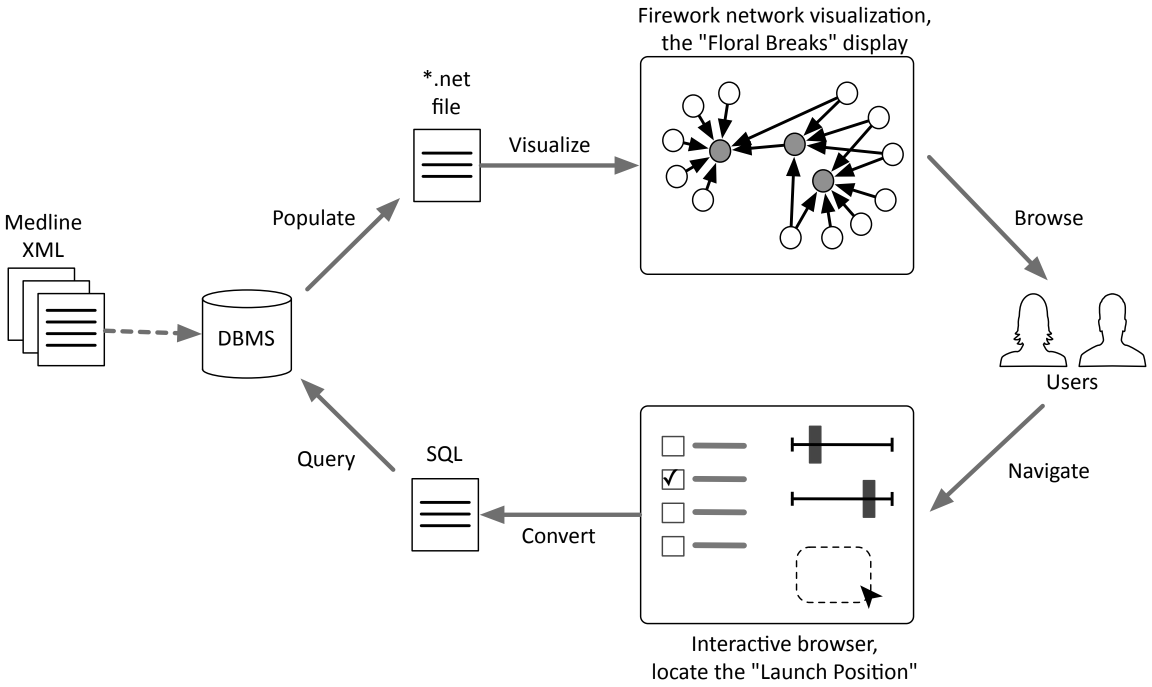


Figure 8: System workflow / information interaction cycle. The XML text data is parsed and stored in a database system. Once a launch position is located, a .net formatted file is extracted from the database system and then bursts one or multiple layers of floral breaks. Users (researchers) browse the network visualization, investigate the aggregated search result, and navigate the space with a control panel. The updated query is converted into structured query language (SQL) and sent to the database system. The system re-populates the network data to start another cycle of search. A dashed line is a one-time process. Solid lines are repeatable.

the $tf_{i,J}$ as follows:

$$\mathrm{tf_{i,J}} = \sum_{k \in J} n_{i,k}$$

where $n_{i,k} = 1$ if the MeSH term i appears in the citation $k \in J$ and $n_{i,k} = 0$ if the term i is absent from the citation k.

We denote the IDF for the term i as idf_i, which is computed as:

$$\mathrm{idf_i} = \log \frac{d}{\sum_{l \in D} n_{i,l}}$$

where d is the number of local citations and the denominator is the frequency count of the term i for all the local citations. We obtain the local MeSH tf-idf by multiplying $tf_{i,J}$ and idf_i:

$$(\mathrm{tf\text{-}idf})_{\mathrm{i,J}} = \mathrm{tf_{i,J}} \times \mathrm{idf_i}$$

For example, a paper with a reference count of 200 would have $d = 200$ where $d \in D$. If a retrieved floral break J has 36 papers, then we have $j = 36$, where j denotes the number of documents in a floral break. If a MeSH term, for example, Transcription Factors, occurs in 30 documents among the retrieved set of 36 papers, $tf_{i,j} = 30$. Assume the term "Transcription Factors" is a common MeSH term and occurs in 160 papers among the 200 local citations. We would obtain $idf_i = \log \frac{200}{160}$. Therefore, the local MeSH tf-idf for the term "Transcription Factors" would be $30 * (\log \frac{200}{160}) = 2.9$. If there exists another MeSH term such as "DNA, Fungal" that also occurs 30 times in the retrieved 36 papers but only occurs 40 times among the total local citations, the local MeSH tf-idf for "DNA, Fungal" would be $30 * (\log \frac{200}{40}) = 20.97$. In this case, we consider "DNA, Fungal" to be a more representative MeSH term than "Transcription Factors" to describe the displayed floral break since it has higher local MeSH tf-idf value.

As shown in Figure 5, the aggregated summary for the five MeSH terms with highest local MeSH tf-idf are different from the five most frequently occurred MeSH terms. We believe that the MeSH term with higher local tf-idf would help identify and reveal the distinct characteristics of a displayed floral break.

5. CONCLUSIONS

In this paper we presented a network analysis approach to studying local citations, exploring the connectivity among papers cited as referenced. We discussed the distribution of in-degree, out-degree, authority, and hub for two selected survey papers. We developed the firework visualization model to display local citation network graphs and to support citation chasing. Our methods facilitate citation chasing as a search strategy and identify overlap in cited paper lists of seed papers. We use local MeSH tf-idf to describe the floral breaks allowing users to make sense of the document collections. We generate long-tail distributions of the document collections so users can gain an overview of the field quickly.

Future work under consideration includes incorporating natural language processing techniques to present the context of the citations – i.e., how a paper was described in other papers. Future work may also apply the firework visualization model at a larger scale, for example, in global citation analysis.

6. ACKNOWLEDGMENTS

This publication was made possible by Grant Number UL1RR024979 from the National Center for Research Resources (NCRR), a part of the National Institutes of Health (NIH). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the CTSA or NIH.

7. REFERENCES

[1] J. Bar-Ilan. Informetrics at the beginning of the 21st century–A review. *Journal of Informetrics*, 2(1):1–52, Jan. 2008.

[2] M. J. Bates. Speculations on browsing, directed searching, and linking in relation to the bradford distribution. In *CoLIS 4 : Fourth International Conference on Conceptions of Library and Information Science: Emerging Frameworks and Methods*, pages 137–150, Seattle, WA, USA, July 2002.

[3] S. C. Bradford. Sources of information on specific subjects. *J. Inf. Sci.*, 10(4):173–180, Oct. 1985. ACM ID: 5035.

[4] R. Delfs, A. Doms, E. Kozlenkov, and M. Schroeder. GoPubMed: Ontology-based literature search applied to GeneOntology and PubMed. *In Proc. of German Bioinformatics Conference. LNBI*, pages 169—178, 2004.

[5] C. Dunne and B. Shneiderman. Improving graph drawing readability by incorporating readability metrics: A software tool for network analysts. Technical Report HCIL-2009-13, University of Maryland, 2009.

[6] D. Ellis and M. Haugan. Modelling the information seeking patterns of engineers and research scientists in an industrial environment. *Journal of Documentation*, 53(4):384–403, 1997.

[7] J. H. Fowler and S. Jeon. The authority of supreme court precedent. *Social Networks*, 30(1):16–30, Jan. 2008.

[8] M. Ghoniem, J. Fekete, and P. Castagliola. A comparison of the readability of graphs using node-link and matrix-based representations. In *IEEE Symposium on Information Visualization*, pages 17–24, 2004.

[9] B. M. Hemminger, D. Lu, K. T. Vaughan, and S. J. Adams. Information seeking behavior of academic scientists. *Journal of the American Society for Information Science and Technology*, 58(14):2205–2225, Dec. 2007.

[10] I. Herman, G. Melancon, and M. Marshall. Graph visualization and navigation in information visualization: A survey. *IEEE Transactions on Visualization and Computer Graphics*, 6(1):24–43, 2000.

[11] R. Hoffmann and A. Valencia. Implementing the iHOP concept for navigation of biomedical literature. *Bioinformatics*, 21(Suppl 2):ii252–ii258, Oct. 2005.

[12] W. W. Hood and C. S. Wilson. The literature of bibliometrics, scientometrics, and informetrics. *Scientometrics*, 52(2):291–314, 2001.

[13] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, Sept. 1999. ACM ID: 324140.

[14] Y. Lin, W. Li, K. Chen, and Y. Liu. A document clustering and ranking system for exploring MEDLINE citations. *Journal of the American Medical Informatics Association*, 14(5):651–661, Sept. 2007.

[15] H. Muller, E. E. Kenny, and P. W. Sternberg. Textpresso: An ontology-based information retrieval and extraction system for biological literature. *PLoS Biology*, 2(11):e309, Nov. 2004. PMID: 15383839.

[16] C. L. Palmer, M. H. Cragin, and T. P. Hogan. Weak information work in scientific discovery. *Information Processing & Management*, 43(3):808–820, May 2007.

[17] A. H. Renear and C. L. Palmer. Strategic reading, ontologies, and the future of scientific publishing. *Science*, 325(5942):828 –832, 2009.

[18] X. Yin, X. Huang, Q. Hu, and Z. Li. Boosting biomedical information retrieval performance through citation graph: An empirical study. In T. Theeramunkong, B. Kijsirikul, N. Cercone, and T. Ho, editors, *Advances in Knowledge Discovery and Data Mining*, volume 5476, pages 949–956. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.