Content Locality in Time-Ordered Document Collections

SILS Technical Report TR-1999-04, September 1999

Charles L. Viles *

School of Information and Library Science

University of North Carolina, Chapel Hill

Chapel Hill, NC 27599-3360, USA

Tel: 1-919-966-5042

Fax: 1-919-962-8071

E-mail: viles@ils.unc.edu

September 7, 1999

Abstract

Using newswire data sources from the TREC corpus, we show that the distribution of relevant documents with respect to time can be decidely non-uniform. Many TREC topics show time-based clustering of relevant documents. We denote this clustering *content locality* and provide a simple metric for its measurement in time-ordered document collections. There is a marked positive correlation between content locality measurements from two time-sychronized data sources. Given this correlation, we show that knowledge of the distribution of content locality in one document source can provide modest improvement in retrieval results in a companion, time-synchronized document source. While this data is preliminary, it illustrates the potential of using time as an additional feature in retrieval.

1 Introduction

The widespread proliferation of on-line newspapers and other time-sensitive or time-ordered document archives has provided the capability for archival searches to a much larger audience than ever before. For example, the on-line version of the *Washington Post* (`http://www.washingtonpost.com/`)

*This work supported in part by DARPA contract N66001-97-C-8542.

currently provides free searches over their entire archive dating back to 1986, but charges fees for articles older than 14 days. The *Chicago Tribune* (`http://www.chicago.tribune.com/`) offers a similar search service for articles from their newspaper dating back to 1985.

Most of these services provide boolean search services with simplistic ranking based on the number of matching search terms. Explicit date range services are usually provided, so if the searcher knows the dates to search, then the query can be limited to that date range. For the searcher who knows that the information they are seeking is inherently topical but is unsure of specific dates, searching is more difficult and the provided tools don't fit.

For these kinds of document databases, there is a clear, fine-grained time-based, ordering of documents. The articles from today's newspaper are "after" the articles in yesterday's newspaper. Since newspapers are inherently topical, we expect to see topics appear as they become newsworthy and disappear as their newsworthiness fades. This kind of temporal locality is evident in one-time events (e.g. natural disasters, terrorist attacks), as well as topics that appear periodically, but on different time scales (e.g. Friday night football games, elections, financial cycles).

Once simple date filtering is done, most information retrieval systems assume that the probability of relevance of a document to a given query is independent of document creation time. Intuitively, this seems like an incorrect assumption, at least for the situations we have just described. In these situations, we'd like to be able to leverage available knowledge of the temporal distribution of a topic in order to improve retrieval.

We call the temporal or spatial "clumping" of topics, *content locality*. Consideration of the spatial dimension can be found in [13, 14]. Here, we are concerned with the temporal dimension. A number of questions arise, including:

- To what extent does locality exist in time-ordered document collections?
- How might locality be measured?
- Most importantly, how might we take advantage of a a topic's temporal locality in order to improve retrieval?

We proceed in Section 2 by providing evidence of the existence and extent of content locality in time-ordered document collections using news feeds from the TREC corpus as the data sources. Results from a simple retrieval experiment illustrate the potential for using content locality information to improve retrieval. We continue in Section 3 with a discussion of the implications of content-locality and how it might be applied in various retrieval scenarios. We finish with a discussion of related work in Section 4 and a summary in Section 5.

2 Existence and Extent of Temporal Locality

As a first step in examining temporal content locality, we visually examined the time-based relevance distributions of TREC [15] queries 51-150 using a 35 month sequence of the AP Newswire. Table 1 contains summary information about this data. Our assumption is that the documents that are relevant to a particular TREC query define a set of documents that are collectively a "topic". [1] In Figure 1 we show two topics, one with high apparent content locality and the other with relatively low content locality.

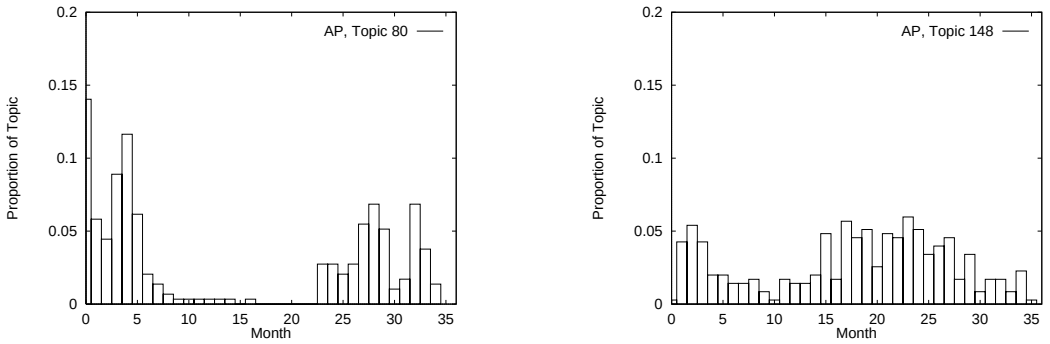


Figure 1: Time-based relevance distributions for two TREC queries. The one on the left shows distinct localization of relevant documents, while the one on the right has a relatively uniform distribution. The topics ask about "1988 presidential platforms" (left) and "war in the Horn of Africa" (right). Data is taken from the AP Newswire, 1988-1990.

Source	Year(s)	Size	Months Represented
AP Newswire	1988 - 1990	242918	35
San Jose Mercury News	1991	90257	12
Wall Street Journal	1991	42652	12

Table 1: Summary information on data sources used in the experiments reported in this paper. All document sets are from the TREC data sets.

[1] In this paper, the term "topic" performs two roles, at some times denoting the statment of a user's information need and at other times referring to the set of related documents that are relevant to that need. The particular meaning should be evident from context.

2.1 A Measure for Locality

While much of the TREC data shows the kind of non-uniformity of relevance exhibited in Figure 1, it is helpful to provide more objective evidence of locality and its extent. To that end, we define a topic-centric measure of locality, σ, that is adapted from a spatial measure proposed in [13] for distributed document collections.

This measure partitions the document collection by site, and then for a particular topic, measures the distance between the actual representation of a topic at a site and the expected representation of the topic at that site in a content-uniform collection. Details follow.

The spatial measure for a topic p at site s is

$$\sigma_{s,p} = (b_{s,p} - c_s)^2 \tag{1}$$

where $b_{s,p} = \frac{n_{s,p}}{n_p}$ is the size $n_{s,p}$ of the topic at the site relative to the overall size n_p of that topic. The coefficient c_s is the expected value of $b_{s,p}$ when content is uniformly distributed throughout the document collection. We expect that this expected value should be directly proportional to the relative size of collection s, so

$$c_s = N_s/N$$

where N_s is the size of collection s and N is the overall size of the document corpus.

Locality for some topic p is denoted σ_p and is determined by summing the locality for that topic over every site and taking the square root. So

$$\sigma_p = \sqrt{\sum_{s=1}^{S} \sigma_{s,p}}$$

and by substitution

$$\sigma_p = \sqrt{\sum_{s=1}^{S} (b_{s,p} - c_s)^2}. \tag{2}$$

Note also that c_s and $b_{s,p}$ are constrained by

$$\sum_{s=1}^{S} c_s = 1$$

and

$$\sum_{s=1}^{S} b_{s,p} = 1$$

since they represent simple proportions of a whole.

The σ measure has the "feel" of a mean-squared error measurement, but does not have an explicit scaling by the number of data points (S in this case). Because of the constraints on c_s

and $b_{s,p}$, it can be shown [13] that $0 \leq \sigma < \sqrt{2}$ in the general case and $0 \leq \sigma \leq 1$ when the size of each sub-collection is equal.

For temporal collections, we partition the entire collection using the time-stamp for the documents to determine which sub-collection a document belongs to. Thus, in a time-ordered collection, a date range defines the temporal equivalent of a "site" in a distributed collection. Once the partition using time-stamps has been accomplished, the interpretation of the above measure for time-ordered collections is obvious and the calculation of σ is straightforward.

There are at least two significant characteristics of temporal collections and the σ measure to remember. First, the granularity of the measure is determined by the width of the date range. The wider the range, the fewer the number of sub-collections in the corpus. Second, there is a definite ordering of each equivalence class in the partition based upon the date range. That is, sub-collection s is quite literally "next to" sub-collection $s+1$ in time. The σ measure does not account for this attribute. If a topic is spread over several contiguous time periods, then a more sophisticated measure is needed to detect this locality and to differentiate it from a simple random ordering in time of the sub-collections that make up the time-ordered archive.

For these reasons, it is best to consider σ as defined above as a gross metric. We use it here to provide a first level ordering of a set of topics based on their calculated temporal locality.

2.2 Locality in the AP Newswire

Using the σ measure, we calculated the temporal locality of the 50 largest topics in the AP newswire portion of the TREC data selected from the 100 TREC topics numbered 51-150. We chose only the 50 largest topics because there are some topics where there are very few relevant documents. For these topics, there is not enough data to accurately measure locality. We chose a time period of one month, which for the almost three year period of the AP documents, corresponds to a partition into 35 sub-collections.

The distribution of locality for the 50 topics is given in Figure 2. Temporal locality as measured by σ varies fairly uniformly from 0.08 to 0.22 with one outlier showing very high locality. Interpretation of the actual value of σ is problematic, as we as yet do not have a way to say that e.g. $\sigma = 0.23$ is low, medium, or high.

As an illustration of the efficacy of the measure in providing a *relative* ordering of topics by their locality, we show the top four topics in Figure 3, and the bottom four topics in Figure 4. A side-by-side comparison of these distributions illustrates the marked difference between topics that are highly localized and topics that are uniformly distributed in time.

In Figure 5 we show the relationship of topic size to locality for topics 51-150 and the AP data. As intutition suggests, there is a general trend toward smaller topics having high locality

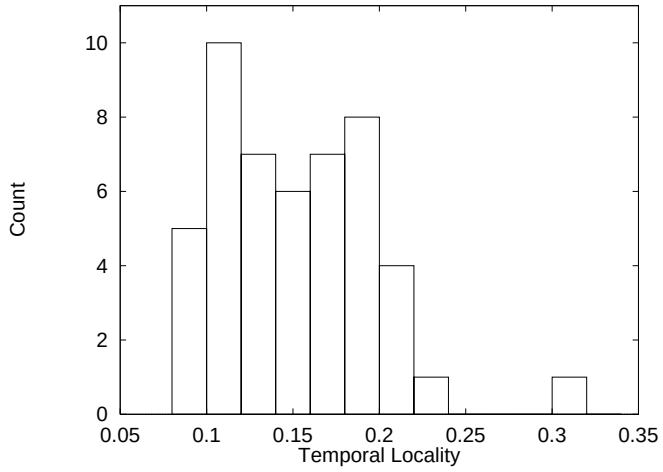


Figure 2: Distribution of locality for the 50 largest topics in the AP Newswire, 1988-1990.

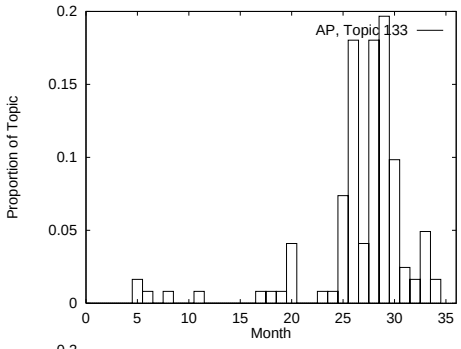
and larger topics having low locality. As mentioned previously, extremely small topics are inherently localized and the measure reflects this characteristic.

2.3 Correlation between Data Sources

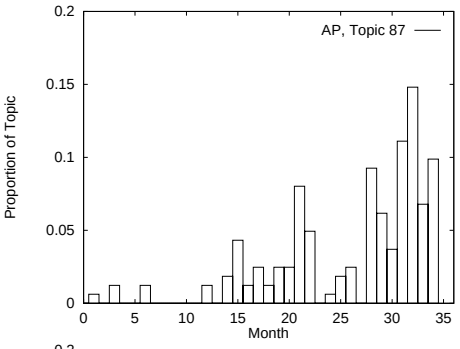
Another interesting avenue to explore is the relation that two document sources covering the same time period have to each other. In Figure 6 we plot σ for the Wall Street Journal (vertical axis) against σ for the San Jose Mercury News (horizontal axis).The figure shows the locality for the 37 common largest topics in WSJ and SJMN.

One simple assumption to make is that if each document source covers the same topic areas in about the same proportion at about the same time, then we would expect that there should be a high correlation between the calculated locality for each source. While Figure 6 shows a positive correlation between the two sources, it is not perfect. One possible explanation is that the simple assumption above may be incorrect for some topics i.e. the papers may not cover the same material in the same proportions. We note also that a high correlation between σ's for two document sources is a necessary but not sufficient condition for determining whether the two are related in time. This is because the calculation of σ does not take into account the position in time of a sub-collection with respect to other sub-collections in the archive.

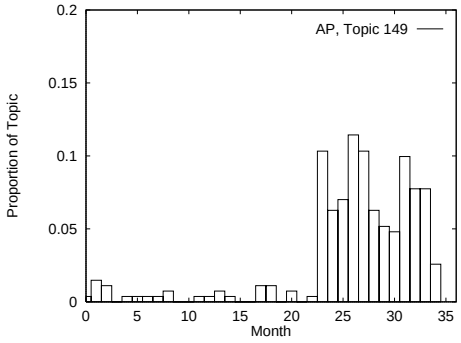
Outliers in Figure 6 are those that show high content locality in one source and low locality in the other. The text of the topics representing the four marked outliers in this figure appears in Table 2. For each topic, there are reasonable explanations for the differing locality. Since the *Wall Street Journal* is devoted to economic and financial topics, it's no surprise that locality for



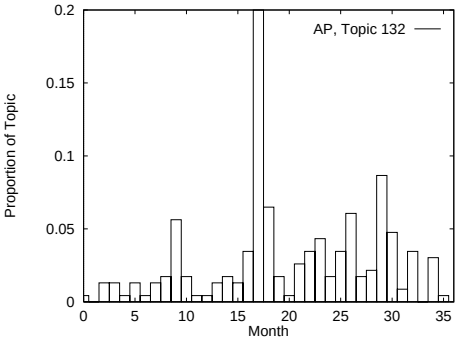
Topic 133: Document will describe some design feature of the Hubble Space Telescope



Topic 87: Document will report on current criminal actions against officers of a failed U.S. banking institution (including all categories of banking).

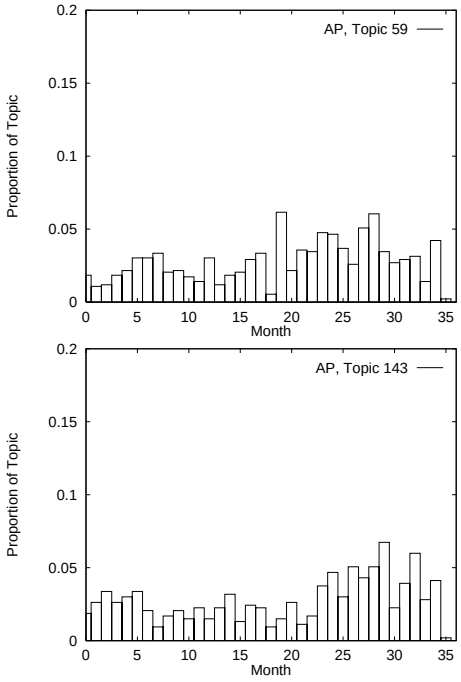


Topic 149: Document will report on specific instances of industrial espionage, including insider trading, or the actions of governments to prevent the theft of economic secrets through legislation, regulation, or law enforcement.



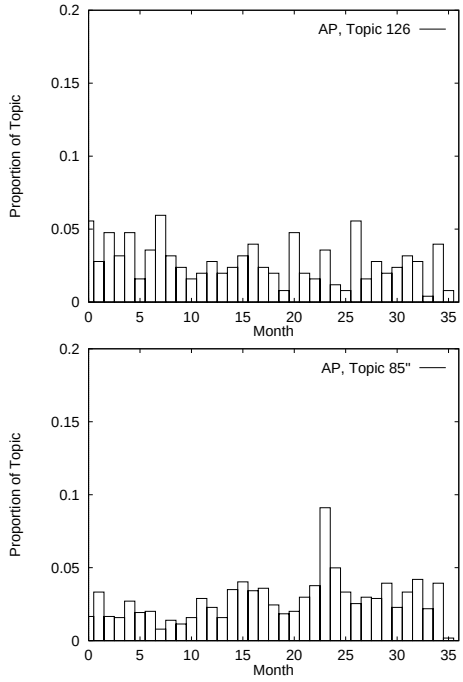
Topic 132: Document will provide cost, technical, and/or performance data on U.S. "stealth" aircraft projects.

Figure 3: TREC topics with the highest calculated locality. Data taken from the AP Newswire 1988-1990 using TREC topics 51-150.



Topic 59: Document will report a type of weather event which has directly caused at least one fatality in some location.

Topic 143: Document will contribute to an analysis of how or at what cost, the U.S. federal government protects, supports, and controls U.S. farming.



Topic 126: Document will discuss ethical issues attendant to contemporary advances in medical technology.

Topic 85: Document will discuss allegations, or measures being taken against, corrupt public officials of any governmental jurisdiction worldwide.

Figure 4: TREC topics with the lowest calculated locality. Data taken from the AP Newswire, 1988-1990 using TREC topics 51-150.

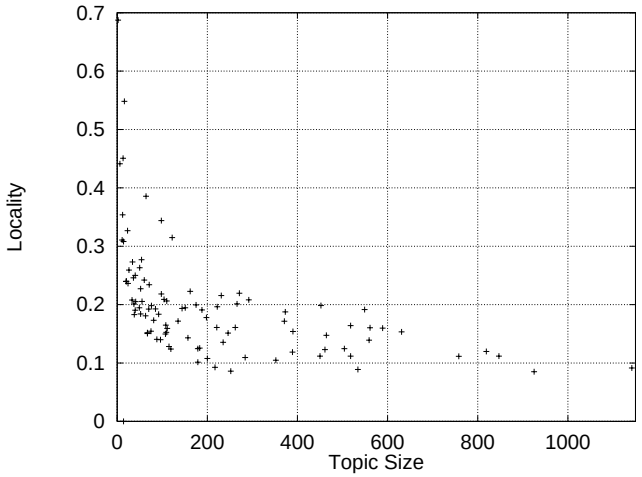


Figure 5: The relationship of temporal locality to topic size. Larger topics tend to have lower locality. Data taken from AP Newswire and TREC topics 51-150.

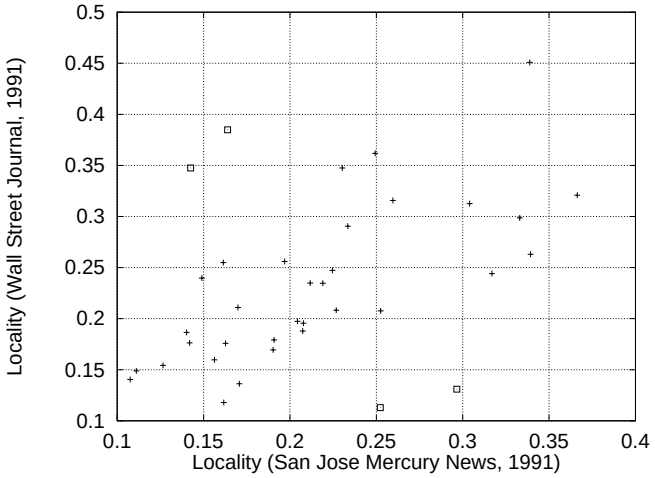


Figure 6: Content locality for two document sources covering the same time period, 1991. Locality for WSJ is on the y-axis and locality for SJMN is on the x-axis. Detail for the four outliers (boxes) is given in the accompanying table.

TopicNum	σ_{SJMN}	σ_{WSJ}	Topic Description
55	0.30	0.13	Document discusses an insider-trading case.
59	0.16	0.38	Document will report a type of weather event which has directly caused at least one fatality in some location.
94	0.14	0.35	Document must identify a crime perpetrated with the aid of a computer.
98	0.25	0.11	Document must identify individuals or organizations which produce fiber optics equipment.

Table 2: Topics whose content locality differs between the WSJ and SJMN sources.

topics related to e.g. "Insider Trading" or "fiber-optic production" is low – these are commonly occuring themes in WSJ. In a more general source like the *San Jose Mercury News* these topics are less common and therefore show higher locality. General news like "weather-related fatalities" is likely to be more common in SJMN and thus exhibit lower locality compared to WSJ. Finally, since SJMN is based in Silicon Valley, "computer-aided crime" is well represented in that source.

The correlation of locality between two document sources suggests that knowledge of the time-base topical distribution in one source might help in retrieval on the other source. We designed a simple experiment to test this hypothesis using the WSJ and SJMN news feeds from 1991.

Source	TREC (WSJ - 1991)
Query Weights	nfx (*idf* variant)
Doc Weights	Lxu (length normalized)
Source for Scaling	TREC(SJMN - 1991)
Query Source	TREC 101-150

Table 3: Details of the retrieval experiment

The experiment was conducted as follows. We selected 50 topic descriptions from the TREC data set [7, 15] (topics 101-150), to run against the WSJ91 data source. Of these 50, 42 had at least one relevant document in WSJ91 and SJMN. Using an implementation of the vector

space model, we ran these 42 topics against the WSJ91 data using length normalized document term weights [12], and an *idf*-based variant for query term weights. We then scaled the resulting similarity scores by the time-based relevance distributions available from SJMN, the data source that is time-synchronized with WSJ91. The scaling operation for the new similarity value $newsim_{d,p}$ for document d and topic p with original similarity $sim_{d,p}$ is

$$newsim_{d,p} = (0.2 * W_{SJMN,date(d),p} * sim_{d,p}) + (0.8 * sim_{d,p})$$

The time-based topical distribution for query/topic p in SJMN is represented by the scaling factor $W_{SJMN,date(d),q}$. For documents that are part of topic p in source r and are dated in time range t,

$$w_{r,t,p} = \quad \text{prob(doc is dated in} \quad t| \quad \text{doc} \quad \in p)$$

and

$$W_{r,t,p} = \frac{w_{r,t,p}}{max_{\forall t}(w_{r,t,p})} \tag{3}$$

The coefficients 0.2 and 0.8 are empirically determined.

The net effect of Equation 3 is to pass unchanged all similarity scores for documents that occur in the most likely time period, and to reduce the scores of all others in proportion to the relative probability for the appropriate time period.

The resulting ranked lists of retrieval results were re-scaled by these new similarity scores and evaluated using standard recall/precision measures. Results are given in Figure 7.

The results from this experiment show a 3-5% absolute improvement in precision at recall levels up to 0.4. Thus, knowledge of the content locality distribution from SJMN provides a small but noticable increase in retrieval quality for high precision searches in WSJ91.

3 Discussion and Implications

It is too early to tell how important the detection and leveraging of content-locality information will be to information retrieval. For queries where exact date information is known (e.g. "June 1989" or "Every October") then detection of locality is not needed since it is explicitly given by the user. In other scenarios however, exposing locality information may help considerably in the retrieval process. The experiment reported here is preliminary and results are modest, but it is apparent that there are good possibilities to improve retrieval.

In the absence of time-based relevance information, the direct measurement of temporal content locality is difficult. As we might expect however, there can be a strong direct relationship between the distribution of a collection statistic like document frequency (the number of documents containing a term) and the distribution of relevant documents. While determination

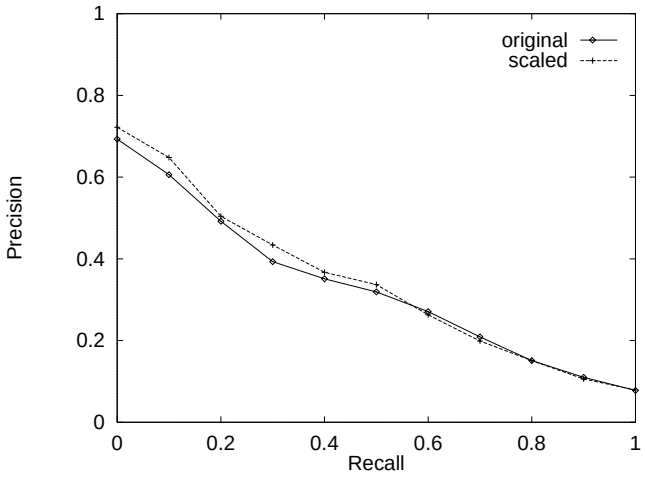


Figure 7: Results from experiment that scales retrieval results("orig") from one document source using topical distributions from a second, time-sychronized source to produce a scaled ranking ("scaled"). The source for scaling information is SJMN and results are from WSJ91.

of the latter is problematic, determination of the former is easy. In Figure 8 we give the weekly tally of document frequencies for three terms in TREC topic 133, which asks about the "hubble space telescope". The relevance distribution for topic 133 is shown in Figure 1. For at least two of these terms, there is a close visual correlation between the document frequency distribution and the relevance distribution. This suggests that document frequency, an easily measurable attribute of any document collection, might be a reasonable indicator of temporal locality.

For adhoc retrieval, one can argue that the retrieval process itself will cause the focusing of search in a particular time period if temporal locality exists for the query in question. In the extreme, if the only documents that contain the query words occur on a particular day, then indirectly you are focusing search on that day. This argument implies that the explicit accounting of temporal information is unnecessary in retrieval because things "take care of themselves". While this is a hypothesis that needs to be tested, we believe that explicitly considering temporal information will help in many cases.

For interactive and feedback related retrieval, one can easily imagine a user providing feedback indicating some initial relevant documents, and then the system concentrating the next search iteration in the temporal neighborhood of the indicated relevant documents.

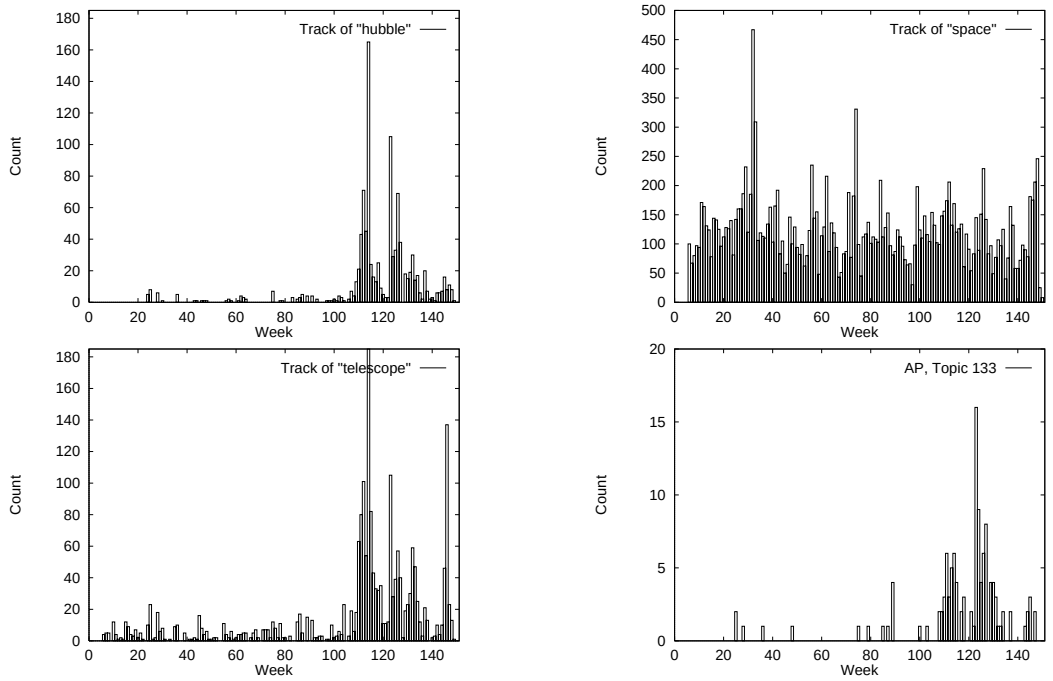


Figure 8: Plots of document frequency for three terms in TREC topic 133. "hubble" (top-left), "space" (top-right), and "telescope" (bottom-left). Relevance distribution for the topic is at bottom-right. Data is from the AP Newswire, 1988-1990.

4 Related Work

Singhal [12] made the observation that in the TREC [15] retrieval environment, longer documents were more likely to be relevant to a query than shorter documents. This observation led to an highly effective length normalization methodology for term weighting that rewards longer documents and penalizes shorter documents.

Length normalization is based on a document characteristic that is not directly relatable to the "aboutness" or topicality of the document, merely an easily determinable feature, its length. It seems reasonable then, to investigate other "non-content" aspects of documents in order to improve retrieval. The particular aspect we consider here is the time ordering of documents in an archive.

The use of relevance information from one source to use for retrieval on another source has elements of classical feedback techniques [10], document filtering [9], and the early TREC routing experiments [5, 6]. The difference in this work is that we use only time-based topical distributions as adjuncts to searching.

In previous work examining spatially-based content locality in distributed document collections [13, 14], two possible metrics were proposed, one based on statistical properties of the constituent document collections and the other based on the non-uniformity of topic distribution in the distributed collection. The one proposed here is based directly on the latter.

Locality as a concept is hardly new. Spatial and temporal locality have been and continue to be integral concepts in a wide variety of areas, including analysis of reference patterns in numerical codes [8], file caching in distributed and networked file systems [11], and caching in various distributed computer systems, World Wide Web servers and browsers being the most obvious current example.

Using document frequency information to focus retrieval in a particular time period has a clear parallel to the collection selection problem in distributed information retrieval, where collection statistics can be used to select a (hopefully) small number of document sources to send a query too [4, 1]. The efficacy of such approaches is a subject of continued study [2].

5 Summary

In this study, we provide compelling evidence of the existence and extent of content locality in time ordered document collections. We provide a metric to measure locality and show that despite some shortcomings, it is effective in providing an ordering of topics by their content locality. We show that knowledge of the time-base content locality in one document collection can improve retrieval in a companion, time-synchronized collection. Our results are clearly

preliminary but are highly encouraging.

References

[1] James P. Callan, Zhihong Lu, and W. Bruce Croft. Searching Distributed Collections with Inference Networks. In *Proceedings of the 18th International Conference on Research and Development in Information Retrieval*, pages 21–29, Seattle, WA, 1995.

[2] James C. French, Allison Powell, Charles L. Viles, Travis Emmitt, and Kevin Prey. Evaluating Database Selection Techniques: A Testbed and Experiment. In *Proceedings of the 21st Conference on Research and Development in Information Retrieval*, Melbourne, Australia, 1998.

[3] Luis Gravano, Hector Garcia-Molina, and Anthony Tomasic. The Effectiveness of GlOSS for the Text Database Discovery Problem. In *SIGMOD94*, pages 126–137, Minneapolis, MN, May 1994.

[4] Donna Harman. Overview of the First Text Retrieval Conference (TREC-1). In *Proceedings of the First Text Retrieval Conference (TREC-1)*, pages 1–20, Gaithersburg, MD, 1992.

[5] Donna Harman. Overview of the Second Text Retrieval Conference (TREC-2). In *Proceedings of the Second Text Retrieval Conference (TREC-2)*, pages 1–20, Gaithersburg, MD, 1993.

[6] Donna Harman. Overview of the Fourth Text Retrieval Conference (TREC-4). In *Proceedings of the Fourth Text Retrieval Conference (TREC-4)*, Gaithersburg, MD, 1995.

[7] Kathryn McKinley and Olivier Temam. A Quantitative Analysis of Loop Nest Locality. In *Proceedings of ASPLOS VII*, pages 94–104, Boston, MA, October 1996.

[8] Douglas W. Oard. The State of the Art in Text Filtering. *User Modeling and User-Adapted Interaction*, 7:141–178, 1997.

[9] Gerard Salton and Chris Buckley. Improving retrieval performance by relevance feedback. *Journal of the American Society of Information Science*, 41:288–297, 1990.

[10] M. Satyanarayanan. Distributed File Systems. In Sape J. Mullender, editor, *Distributed Systems*. ACM Press, 1989.

[11] Amit Singhal, Chris Buckley, and Mandar Mitra. Pivoted Document Length Normalization. In *Proceedings of the 19th International Conference on Research and Development in Information Retrieval*, pages 21–29, Zurich, Switzerland, August 1996.

[12] Charles L. Viles. *Maintaining Retrieval Effectiveness in Distributed, Dynamic Information Retrieval Systems*. PhD thesis, University of Virginia, 1996.

[13] Charles L. Viles and James C. French. Content Locality in Distributed Digital Libraries. *Information Processing and Management*, 35(3):317–336, 1999.

[14] Ellen Voorhees and Donna Harman. Overview of the Fifth Text Retrieval Conference (TREC-5). In *Proceedings of the Fifth Text Retrieval Conference (TREC-5)*, pages 1–28, Gaithersburg, MD, 1996.