Dependence of Binary Associations on Co-occurrence Granularity in News Documents

Krishnaprasad Thirunarayan[1] , Trivikram Immaneni[1], and Mastan Vali Shaik[1]
[1]Department of Computer Science and Engineering
Wright State University, Dayton, OH-45435, U.S.A.

Abstract *- We describe and formalize an approach to correlate binary associations (such as between entities and events, between persons and events, etc.) implied by News documents on the co-occurrence granularity (such as document-level, paragraph-level, sentence-level, etc.) of the corresponding text phrases in the documents. Specifically, we present both qualitative and quantitative characterization of searching News documents: former in terms of the nature of the content and the queries, and latter in terms of a metric obtained by adapting the notions of precision and recall. Specifically, the approach tries to reduce the manual effort required to analyze the News documents to compare the three alternatives for granularity of co-occurrence. Furthermore, the analysis suggests ways to improve retrieval performance as illustrated by applying our findings to News documents for the year 2005.*

Keywords*:* News documents, index and search, timelines, precision and recall, binary associations, co-occurrence granularity.

1 Introduction

Search engines for News documents use metadata in the form of SmartIndex terms, in addition to full-text search, for scalability and effectiveness [1]. SmartIndex terms and their weights are obtained by analyzing the News document that takes into account the frequency of occurrence of a phrase in the text, its position (in headline, in lead paragraph, etc), amount of user interest in the topic, etc. From the perspective of indexing and search, SmartIndex terms provide a primitive ontology for capturing significant aspects of News content. Specifically, the metadata (SmartIndex terms) associated with a document abstracts and normalizes content. For instance, to retrieve documents about the entity "Microsoft" and the event "Mergers and Acquisition", the search engine uses an index based on these concepts.

An association timeline (such as entity-event timeline, person-entity timeline, country-event timeline, etc) is a graph of the number of News documents supporting the association, versus the dates. A timeline-based interface can be exploited for searching, visualizing, and accessing the answer sets of News documents obtained in response to a query, which is the basis for building our timeline application. See Yahoo! Finance [2], Google Trends [3], etc for other practical examples. In particular, Google Trends is a timeline which tries to determine relative interest in a topic on the basis of the number of searches on the topic (Search volume graph) and the number of News stories involving the topic (News reference volume graph) over time [4, 5]. Other related work that use the time dimension in the context of analyzing News documents can be found under Temporal Information Retrieval in Alonso and Gertz [6], New Event Detection by monitoring chronologically-ordered News streams in Giridhar and Allan [7], and a framework for developing personalized News Search Engine that ranks News both on the source and "freshness" in Gulli [8].

For concreteness, we focus on entity-event timelines though our timeline application has been generalized to various binary associations over the following set {entity, event, person, country}. A News document supports an entity-event association if it contains document-level metadata corresponding to the entity and the event. The use of document-level metadata scales well in practice and has been a de facto standard. However, the notion of support based solely on document-level metadata is unsound when a News document contains multiple News stories, or when a story involves several different associations. This is due to the "crosstalk" caused by pairing an entity from one story with an event from another story.

To eliminate mis-associations, we explore two separate refinements – one that redefines support in terms of co-occurrence of entity-event pair in a paragraph, and another that redefines support in terms of co-occurrence of the entity-event pair in a sentence, instead of their mere appearance in the document. Similar ideas have been explored recently by Jin and Srihari [9] in developing graph-based models for text for the purposes of ranking concept chains, where they represent and infer the strength of association among concepts using concept co-occurrence relationship. Popov et al [10] use frequency of occurrence to further grade semantic associations. In our timeline application, due to co-references (such as pronouns, aliases, synonyms, acronyms, etc), this refinement of the criteria for support (that is, going from document-level co-occurrence to sentence/paragraph-level co-occurrence) promises to improve precision at the expense of recall. However, it is not obvious where to draw a line to rationalize this tradeoff as explained in Section 2.

In Section 3, we illustrate the dependence of binary association on co-occurrence granularity and shed light on the idiosyncratic nature of News documents. These observations lead us to the fundamental research problem: How can we determine, quantify, and improve the reliability of inferred associations in a scalable way? In Section 4, we formalize a quality criterion for choosing appropriate level of co-occurrence for a document cluster by adapting traditional notions of precision and recall in a novel way [11]. In effect, we attempt to quantify which granularity level to employ for a given query. This enables better understanding of the nature of the News dataset, and sheds light on situations where each criterion has its merit. Eventually, we can apply these results to determine which refinements are promising for a given News document dataset and verbalize the characteristics of the queries for which the refinement is effective. Section 4 also shows a fragment of the experimental results obtained for News documents for the year 2005. Section 5 concludes with suggestions for future work.

2 Relationship to Information Retrieval

Data retrieval aims at determining all objects that satisfy a semantically well-defined query, while information retrieval aims to decipher user information need and interpret document contents in order to satisfy a query. Given that natural language text (News document) can be ambiguous, and user information need cannot always be articulated satisfactorily in a provided query language, IR systems are typically evaluated with respect to standard benchmarks in terms of precision and recall (and its variants) [12]. In other words, for "testing" an IR system, the semantics of the document contents and the query is specified *indirectly* using the benchmark. In the case of News documents, there are no standard benchmarks available that specifies the binary associations implied by its documents that meet our needs. Instead, in Section 3, we provide illustrative examples of pros and cons of document-level metadata for inferring binary association, and argue for considering paragraph-level metadata and sentence-level metadata. Specifically, in Section 3, we show that the use of paragraph-level co-occurrence and sentence-level occurrence can improve the reliability of certain query answer sets but not all. In Section 4, we formalize criteria for determining when paragraph-level (resp. document-level) co-occurrence should be used in preference to document-level or sentence-level (resp. sentence-level or paragraph-level) occurrence, and vice versa, for inferring binary associations. Thus, Section 4 provides a novel evaluation metric for our approach that can be used on a standard benchmark (when available). Furthermore, our approach tries to reduce the manual effort required to analyze the News documents for comparing the three alternatives for granularity of co-occurrence.

3 Baseline Timeline System and its Enhancement: An Analysis

The entity-event timeline application is driven by a query involving entity, events, and a duration of interest. A point on an entity-event timeline gives the number of News documents relevant to the entity-event pair on the corresponding date. One can also obtain document headlines, document content, and the document cluster label from this timeline interface. See Figure 1 on the last page. The timeline generation uses an entity-based index that has already been computed offline. An important aspect of the baseline system is that all the analysis for timeline generation is based on the document-level metadata available explicitly in the News documents.

Determining associations between entities and events based on the document-level (global) metadata can result in mis-associations. To improve the reliability of timeline, we explored more fine-grained paragraph-level (resp. sentence-level) support for associations. An entity-event pair has paragraph-level (resp. sentence-level) support if the entity-event pair is referenced within a paragraph (resp. sentence), that is, there are words/phrases denoting the entity and the event. Note that this determination is non-trivial in practice, and requires proprietary extension of named entity and event recognition software for content analysis.

To better understand the nature of News documents and the effect of using different co-occurrence granularity as criteria for inferring (entity-event) associations, consider the following examples (which intentionally contain some idiosyncratic aspects of News document content and the associated metadata):

Document-1: …

* When Kellogg Co. announced plans to bring back $1 billion in foreign profits to the United States this year thanks to new corporate tax breaks, it said it would use those funds to do such things as develop new products and explore potential acquisitions. It didn't mention directing that money toward creating jobs. * …

Document-2: …

* Rexam eyes expansion in Eastern Europe and Asia *

* REXAM PLC, which produces a billion cans a week for the likes of Coca-Cola and Anheuser-Busch, is looking for bolt-on acquisitions to grow in Eastern Europe and Asia after recent success in Russia. Underlying pre-tax profits rose 26% to £300 million on a 4% increase in sales from ongoing operations to £3.12 billion. *

* Sankyo and Daiichi confirm £4bn merger *

* *** News Story *** *

* VT radars in on £30m Royal Navy contract *

* *** News Story *** *

* Medisys axes boss after £6.8m assets writedown *

* *** News Story *** *

…

• Document-1 associates KELLOGG CO with JOB CREATION in spite of the presence of negation. This is done by inspecting consecutive sentences within a paragraph, effectively resolving the co-referent "it". Note that the search engine does not comprehend the content in the natural language processing sense.

• Document-2 is a classic example of a document with multiple News stories that can give rise to mis-associations using document-level metadata. Specifically, document-level metadata such as COMPUTER & ELECTRONICS MFG, COMPUTER MAKERS, DEFENSE CONTRACTING, NAVIES, etc are mis-associated with REXAM PLC. The entity is part of the first story, while the "events" arise from the second and the third stories. Observe that these mis-associations can be eliminated by using paragraph-level analysis.

• Paragraph-level association of metadata such as COMPANY PROFITS with ANHEUSER-BUSCH COS INC can be eliminated by going to sentence-level. However, the net effect is mixed because this refinement is correct with respect to ANHEUSER-BUSCH COS INC but incorrect with respect to REXAM PLC. This shows that, in general, each granularity-level has its pros and cons.

• The sentence-level association of MERGERS & ACQUISITIONS with ANHEUSER-BUSCH COS INC and REXAM PLC is also tricky. In reality, only the association of MERGERS & ACQUISITIONS with REXAM PLC is sound, and this can only be ascertained using clausal-level analysis.

• Pragmatically, document-level analysis is efficient, while paragraph-level and sentence-level analysis, which can improve precision, is compute-intensive. This motivates the need for novel indexing structures.

To understand the relationship between the metadata (such as OIL & GAS PRICES, DISEASES & DISORDERS)), the relative locations of the text phrases corresponding to them (such as fuel costs, heart attacks) and the overall document content, we generated triples of numbers of supporting documents (nd, np, ns) (corresponding to document-level, paragraph-level and sentence-level support), for every "significant" entity-event pair, for News document for the entire year 2005 (150 GB). Table 1 (on the last page) shows a number of entity-event queries and the number of documents retrieved based on document-level, paragraph-level, and sentence-level association. (This table shows a few rows of a table generated from a small fragment of the dataset.) This illustrates how different criteria affect the answer set. In general, more stringent criterion improves precision and reduces recall. Specifically, in Section 4, we will describe how to interpret these numbers to determine which criterion to use for improving reliability of the timeline. That is, when is it reasonable to stick with document-level support criterion in the interest of efficiency, and when is it desirable to expend extra effort involved in computing paragraph-level or sentence-level for reliability.

The analysis of such tables and the corresponding documents yielded a number of idiosyncratic examples of mis-associations and their causes.

(a) There are documents with document-level metadata that does not seem to be the subject of the document. For instance, a document may carry an ABC NEWS tag not because the document is about the News reporting agency, but because the document is a TV News broadcast transcript containing reference to the reporter's affiliation such as "Diane Sawyer of ABC News …", etc.

Another example is the redundant occurrence of stock exchange names (NYSE, AMEX, NASDAQ) in a document about stocks. Sentence-level analysis and special treatment of publisher tags can remedy this problem. (*Publisher/ Stock Exchange Tags Case*)

(b) There are documents for which entity-event associations derived from document-level entity-event tags are unsound but paragraph-level entity-event phrase co-occurrence are sound. This is especially the case when a document contains fragments of multiple independent short News stories in print or multiple TV News stories (such as about Inaugural function for Pope Benedict XVI followed by Martha Stewart's house arrest). However, moving to sentence-level co-occurrence may cause incompleteness in the presence of aliases (Cardinal Ratzinger vs Pope Benedict XVI vs Holy Father, or GM Cars vs Chevrolet/Cadillac). (*Multiple Independent Stories*)

(c) There are documents describing stock prices and changes in stock quotes (with headlines such as "Big movers on the stock market") that involve companies (such as IBM, GE, Eli Lilly, Citigroup, GM, etc) from different market sectors (such as Hardware, Pharmaceuticals, Automotive, etc) and fluctuations due to different reasons (such as Oil Prices, Litigation, etc). This can potentially be a rich source of mis-associations. Derivation of entity-event associations solely on the basis of sentence-level entity-event co-occurrence can improve precision in these cases. (*Stock News documents*)

(d) There are some broad Question-Answer documents whose document-level tags can lead to mis-associations. The precision can be improved by using paragraph-level or sentence-level analysis. (*Q&A documents*)

(e) A sentence containing "Kate Snow of ABC News" gets erroneously tagged with WEATHER and ABC News showing that inferred sentence-level associations can be unsound in general. (*Content-based Errors*)

To summarize, this analysis shows that we can potentially characterize News documents in terms of their content, and then choose appropriate co-occurrence granularity to infer associations, as determined by the approach in Section 4.

4 Quantifying Timeline Improvement over the Baseline

In order to uncover an optimal granularity for co-occurrence of entity-event phrases for determining document support for an entity-event query, we propose criteria based on a variant of the traditional notions of precision and recall. Specifically, to establish paragraph-level as the appropriate granularity for co-occurrence when computing entity-event query results, we need to establish that going from document-level to paragraph-level shows far greater increase in precision compared to reduction in recall, while going from paragraph-level to sentence-level diminishes recall much more than it improves precision.

We focus on the manual inspection of documents that are eliminated when going from document-level (resp. paragraph-level) to paragraph-level (resp. sentence-level), to determine estimates for changes in precision and recall. Our approach tries to minimize the manual effort required to analyze the News documents to compare the three alternatives for granularity of co-occurrence. This is particularly significant for evaluation in the absence of a standard benchmark.

Let nd, np, and ns be the number of documents that support an entity-event association at document-level, paragraph-level, and sentence-level respectively. We assume that the recall for an entity-event query on the basis of the document-level existence of the corresponding metadata is 100%, and the precision for an entity-event query on the basis of the sentence-level co-occurrence of the corresponding phrases is 100%.

Let the documents (nd – np) containing entity-event phrases in different paragraphs be partitioned into two groups: those that are found relevant to the entity-event query (say of size n_rP) and those that are found irrelevant to the entity-event query (say of size n_iP). [nd – np = n_rP + n_iP] (Observe that relevancy is based on human judgment and only (nd – np) need to be manually analyzed.) Then the "relative" precision at the document level is (nd – n_iP) / nd, or equivalently, (np + n_rP) / nd, assuming optimistically that np documents are relevant. In other words, if there are very few irrelevant documents in which entity-event phrases occur in different paragraphs, then the precision at document-level will be high. The "relative" recall at the paragraph-level is np / (np + n_rP). In other words, if there are many relevant documents in which entity-event phrases occur in different paragraphs, then the recall at paragraph-level will be low. So, to prefer paragraph-level query results over document-level query results, the overall improvement in precision should offset the overall reduction in recall. Thus, we can define a quality factor QF_pd for each entity-event query (or for that matter any document set) by adapting [Recall at paragraph-level / Precision at document-level] as [(np+1) * (nd+1) / (np+1+n_gP)) ^ 2]. (The increment is to avoid any potential unpleasantness when the parameter becomes 0.)

A similar analysis can be carried out to determine when we should prefer sentence-level associations over paragraph-level associations as follows: Let the documents (np – ns) containing entity-event phrases in different sentences be partitioned into two groups: those that are found relevant to the entity-event query (say of size n_rS) and those that are found irrelevant to the entity-event query (say of size n_iS). The precision at the paragraph-level is (np – n_iS) / np, or equivalently, (ns + n_rS) / np, assuming optimistically that ns documents are relevant. The recall at sentence-level is ns / (ns + n_rS). So, to prefer sentence-level query results over paragraph-level query results, we can define a quality factor QF_sp for each entity-event query (or for that matter any document set) by adapting [Recall at sentence-level / Precision at paragraph-level] as [(ns+1) * (np+1) / (ns+n_rS+1) ^ 2]. Note that, if (n_rS >> n_iS), then paragraph-level granularity is preferred over sentence-level granularity. (We can also redefine sentence-level quality factor to understand the cumulative effect of refining from document-level to sentence-level as QF_sd = [(ns+1) * (nd+1) / (ns+n_rS+n_rP+1) ^ 2].)

Through experiments and manual analysis of document clusters over a wide range of queries, we can determine the various quality factors, to determine the most effective granularity of co-occurrence. To summarize: If, for a document cluster obtained as a query result, QF_pd >> QF_sp, then there is benefit in going to paragraph-level. Similarly, if QF_sp >> QF_pd, then there is further benefit in going to sentence-level.

The above analysis provides quantitative support to our qualitative observations and intuitions about the benefits of more focused analysis in the context of querying News documents. Table 2 (on the last page) shows that queries involving REXAM PLC – NAVIES, GOOGLE – OIL & GAS PRICES, and KELLOGG CO – EMPLOYMENT GROWTH, etc benefit from paragraph-level co-occurrence criteria, while the J C PENNEY CO INC – OIL & GAS PRICES query can benefit from sentence-level co-occurrence

criteria. This follows from the fact that for the former queries the QF_pd value is high, while for the latter query the QF_sp value is high. For REXAM PLC – NAVIES query, the result set contains documents with multiple News stories that are removed by paragraph-level analysis. For GOOGLE – OIL & GAS PRICES query, the only document that survives even sentence-level analysis contains an indirect contextual association of "current prices" with OIL & GAS PRICES via latter's connection to Indonesia's economy, and the relationship between Indonesia's net worth to Google's net worth. KELLOGG CO – EMPLOYMENT GROWTH query benefits from paragraph-level analysis that effectively infers association through co-reference. The sentence-level analysis benefits J C PENNEY CO INC – OIL & GAS PRICES query by eliminating association with OIL & GAS PRICES, thereby improving precision. However, the case of J C PENNEY CO INC – COMPANY EARNINGS query is idiosyncratic. "Penney" is not recognized as a reference to the entity J C. PENNEY by the proprietary named entity recognizer, thereby reducing recall (shown in RED in Table 2). If we were to change the document by updating "Penney" to "J. C. Penney", we reach the conclusion that document-level analysis is adequate (shown in GREEN in Table 2).

Similar analysis done on the table generated for a larger dataset shows that precision of the results of an entity-event query can be improved by considering document-level co-occurrence of entity-event metadata for general documents, paragraph-level co-occurrence of entity-event phrases for multiple News item documents and Questions-Answers documents, and sentence-level co-occurrence of entity-event phrases for Stocks related documents, consistent with our qualitative observations and intuitions.

5 Conclusions and Future Work

The work reported here arose in the context of improving the reliability of our Timeline Application used for searching and visualizing News document datasets [4]. We built a timeline application and a number of analysis tools to study empirically the relationship between the distribution of entity-event phrases with respect to document/paragraph/sentence boundaries, and used it to develop heuristics to determine associations supported by document content.

We determined, qualitatively, examples of "problematic" documents that deserve finer co-occurrence granularity, and have sketched an approach to quantify these results. We also used a subset of documents from a 150GB News dataset for the year 2005 to validate our formalization. We are still looking for a suitable benchmark and for analysts who can independently judge documents, to carry out more thorough evaluation. We are exploring ways to generate document sub-clusters for which the effort involved in computing the co-occurrence at paragraph-level or at sentence-level will be offset by substantially improved precision. Eventually, we propose to characterize such documents clusters in terms of local features to pave way for scalable, practical, and online recognition algorithms. This will enable us to determine the granularity of co-occurrence of entity-event phrases that can be used for constructing answer sets as a function of the characteristics of the query and of the document dataset with reasonable trade-off between precision and recall. In the near future, we are developing and implementing a suitable metadata-based indexing scheme that makes explicit co-occurrence granularity to compute refined timelines efficiently.

Acknowledgments. We are indebted to Don Loritz for enlightening discussions and assistance throughout this project.

6 References

[1]. http://www.lexisnexis.com
[2]. http://finance.yahoo.com/
[3]. http://www.google.com/trends/
[4]. K. Thirunarayan, T. Immaneni, and M. Shaik, "Selecting labels for news document clusters", In: Proceedings of 12th International Conference on Applications of Natural Language to Information Systems (NLDB 2007), LNCS 4592, pp. 119-130 (2007)
[5]. K. Thirunarayan, and T. Immaneni, "A coherent query language for XML", In: Journal of Intelligent Information Systems. (2008) (to appear)
[6]. O. Alonso and M. Gertz, "Clustering of search results using temporal attributes", Proceedings of 29th ACM SIGIR Conference, pp. 597-598 (2006).
[7]. K. Giridhar and J. Allan, "Text classification and named entities for new event detection", Proceedings of the 27th Annual International ACM SIGIR Conference, New York, NY, USA. ACM Press., pp. 297—304 (2004)
[8]. A. Gulli: "The anatomy of a news search engine", Proceedings of 14th International World Wide Web Conference, 880-881 (2005).
[9]. W. Jin and R. K. Srihari, "Graph-based text representation and knowledge discovery", Proc. ACM Symposium on Applied Computing (SAC), Seoul, S. Korea, pp. 807-811 (2007)
[10]. Borislav Popov, Ilian Kitchukov, Krasimir Angelov, Atanas Kiryakov: "Co-occurrence and ranking of entities", Ontotext Technology White Paper (2006)
[11]. Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press (2008)
[12]. Ricardo Baeza-Yates and Ribeiro-Neto. Modern Information Retrieval. ACM Press Series/Addison Wesley (1999)

Figure1: Web-based Timeline application implemented in Java and Adobe Flex

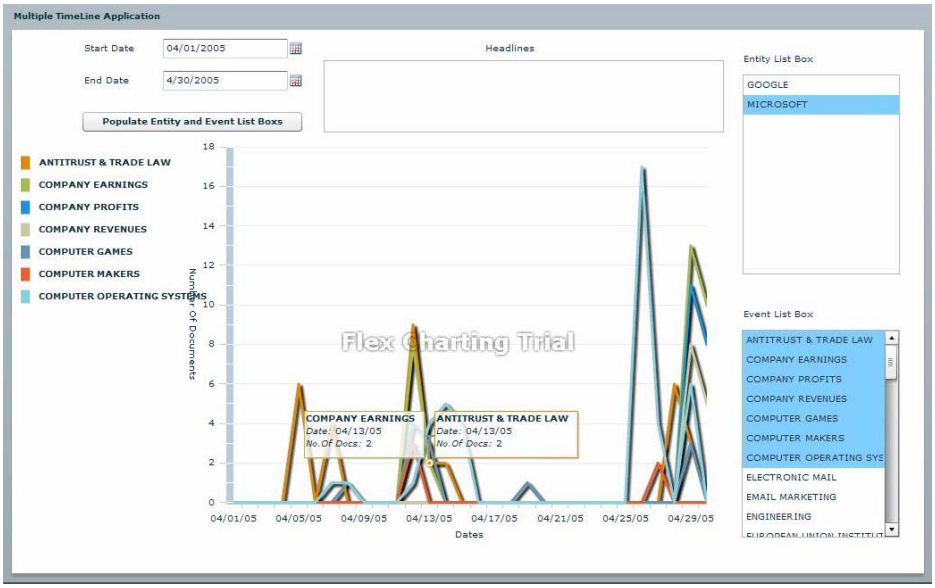

Table 1. Refinement of association using different co-occurrence criteria

Entity	Event	#Doc	#Para	#Sent
WILLIAM WRIGLEY JR CO	MERGERS & ACQUISITIONS	5	5	0
NORTHWEST AIRLINES CORP	OIL & GAS PRICES	8	3	3
AMERICAN BROADCASTING COS INC	TERRORISM	10	6	5
J C PENNEY CO INC	BUYINS & BUYOUTS	8	8	0
R R DONNELLEY & SONS CO	MERGERS & ACQUISITIONS	9	9	0
US DEPARTMENT OF ENERGY	FUEL CELL TECHNOLOGY	13	13	0
J P MORGAN CHASE & CO	MERGERS & ACQUISITIONS	9	9	0
R J REYNOLDS TOBACCO CO	SUITS & CLAIMS	6	6	0
VOLKSWAGEN AG	JUSTICE DEPARTMENTS	9	5	5
REXAM PLC	NAVIES	5	0	0

Table 2. Quality Factor: Quantifying different co-occurrence criteria

Entity	Event	nd	np	ns	QF_pd	QF_sp
REXAM PLC	NAVIES	5	0	0	6	1
GOOGLE INC	OIL & GAS PRICES	6	1	1	3.5	1
OAO GAZPROM	BANKRUPTCY COURTS	11	8	7	1.33	1.125
KELLOGG CO	EMPLOYMENT GROWTH	4	4	0	1	0.2
J C PENNEY CO INC	**COMPANY EARNINGS**	**3**	**3**	**0**	**1**	**0.18**
J C PENNEY CO INC	OIL & GAS PRICES	3	3	0	1	4
J C PENNEY CO INC	*COMPANY EARNINGS*	*3*	*3*	*3*	*1*	*1*