CDIP: *C*ollection-*D*riven, yet *I*ndividuality-*P*reserving Automated Blog Tagging*

Jong Wook Kim
Comp. Sci. and Eng. Dept.
Arizona State University
Tempe, AZ 85287, USA
jong@asu.edu

K. Selçuk Candan
Comp. Sci. and Eng. Dept.
Arizona State University
Tempe, AZ 85287, USA
candan@asu.edu

Junichi Tatemura
NEC Labs, America
10080 Wolfe Rd,
Cupertino, CA, 95014, USA
tatemura@sv.nec-labs.com

Abstract

With the success of blogs as popular information sharing media, searches on blogs have become popular. In the blogosphere, tagging is used as a means of annotating blog entries with contextually meaningful keywords, which enable users more easily locate blog content. Yet, although tags provided by bloggers are effective for organizing blog entries, in many cases, they are not always sufficient in properly capturing the semantics of the blog content. In our previous work [7], we observed that there exists large degree of content overlap (not only in the form of quotation/commentary pairs, but also as content borrowing across media outlets) among blog entries, which makes it hard for effective, discriminating keyword searches. In this paper, we further note that these implicit or explicit quotations could be leveraged to identify the contexts in which entries occur; thus, resulting in more effective tagging. Thus, we propose CDIP (a collection-driven, yet individuality-preserving tagging system) which relies on relationships provided by quotation/reuse detection and semantic-focus analysis to automatically tag the blogs in such a way that, not-only the related blogs share tags, but also individuality of the entries is preserved for discriminating tag-based accesses.

1. Introduction

The *blogosphere* is doubling in size about once every 5 months and about 30-40,000 new weblogs are being created each day [1]. As blogosphere is growing, it is becoming more and more difficult for individuals to navigate through it. Especially, in many cases, the hyperlinks in blogosphere fail to capture the topical relevance within contents because they are often created by the mutual awareness of the bloggers (for instance, friend relationship) rather than topical closeness in contents. Thus, the hyperlink-based browsing that mirrors the popular access mechanisms for the general web cannot be successfully applied to the blog entries.

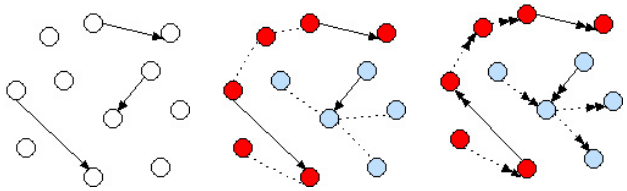


Figure 1. Proposed approach: (a) a number of blog entries and the hyperlinks between them, (b) reuse and quotation links established after the quotation analysis, and (c) the focus directions.

Another way for accessing blogs is a tag-based search. Tag provides an easy way to search and index blog entries through annotating them with short textual words. The idea of tagging is not a new concept: it has been used in the various areas such as photo-sharing applications in the web. Recently, with the quick growth of the blogsphere, tag-based searches have been introduced to the blogsphere to overcome the shortage of the hyperlink-based browsing. However, since current tag-based searches are based on the tags provided by blogers, the search results are vulnerable to blog spams that are maliciously created to increase the search engine ranking, and consequently can not satisfy the users' expectation. Thus, we argue that blog entries necessitate alternative tagging strategies, which leverages *semantic* relationship provides by the contents of blog entries.

To address this shortage, in this paper, we develop a tagging method that relies on the keyword propagation technique. In our previous works [7], we observed that in the blogsphere, *direct* (i.e., quotations as well as *content borrowing* across media outlets) and *indirect* (i.e., tracking the same real-world event from the same information sources) content re-use are significantly more prominent in blogs than general web pages. Based on this observation, our key-

*This work was done at NEC Labs, America

word propagation and tagging exploits the *semantic relationship* provided by the content-reuse and the relative contents of the two blog entries to propagate keywords to enrich the individual entries. In particular, context-aware semantic keyword propagation and tagging is especially important in blogosphere, where entries are more contextual than other document/web search.

1.1 Problem Statement: Collection-Driven, yet Individuality-Preserving Tagging of Blog Entries

Our main goal is to enable the user to perform effective tag-based searches and navigation in the blogosphere. In this process, we rely on the fact that large degrees of content overlaps in the form of quotation/commentary pairs and content borrowing across media outlets are common. In particular, we use content overlaps across blog entries as conduits that guide keyword and tag propagation. However, we highlight that this process should associate each blog entry with *keyword* and *tag* vectors that are not only *contextually-informed*, but also *discriminating*:

- *contextually-informed* tag vectors would reflect not only the content of the entry, but also the context in which this entry is presented.
- *discriminating* vectors would be able to differentiate individual entries among the crowd of related content.

These two goals, on the other hand, may conflict. *Unless keyword- and tag-vectors of individual blog entries are treated carefully*, propagated keywords and tags could render these entries less and less indistinguishable. Considering that blog entries already contain large degrees of content overlaps, this might be detrimental to any search or mining system that relies on these keywords.

In this paper, we highlight that although the task of obtaining *contextually-informed*, but also *discriminating* tag vectors is challenging, it is not unsurmountable. In particular, we observe that understanding how specifically two given blog entries relate to each other can go a long way toward establishing the context in which these entries should be considered. Furthermore, a precise understanding of the relationship between the two entries can prevent irrelevant keywords being *exchanged* between them. Thus, we describe the properties of the *keyword and tag exchange* between blog entries in two desiderata:

Desideratum 1.1 (Relatedness) *Keyword and tag exchange between blog entries should be performed between content that relate to each other not only explicitly (through hyperlinks) but also implicitly (through content reuse and overlaps).*

In other words, before tag propagation, the system should place each blog entry into implicit or explicit "quotation" structures.

Desideratum 1.2 (Discriminating) *Furthermore, while contextually enriching the blog entries, any keyword or tag exchange between entries should at least preserve the original entries' discriminating power.*

Although measuring the relative importance of a quotation between two entries can help preventing the exchange of completely irrelevant keyword and tags, we highlight that understanding the *semantic-focus* of the entries will further improve the preservation of the discriminating power of the entries. Intuitively, while a more general entry is expected to subsume the content of related entries, the topics covered by a focused entry is expected to be more specialized. Thus, understanding the semantic relationships between two blog entries can help preserving the discriminating power of the entries after propagation.

Figure 1 provides an overview of the proposed approach for content enrichment of blog entries:

(a) an initial set of blog entries are identified (possibly through an initial keyword search),
(b) quotation analysis helps establish specific relationships between the various entries in the given set, and
(c) further analysis helps establish the *semantic-focus* directions between highly related entries.

Finally, this information is used for *collection-driven, yet individuality-preserving* tagging of these blog entries.

The structure of the paper is as follows. Next section presents the related work. Section 3 briefly describes the data representation. Section 4 discusses how the proposed approach identifies re-use, sharing, and relatedness across blogs. Section 5 describes the focus analysis and Section 7 discusses the discrimination preserving content-enrichment process. Section 8 experimentally verifies the performance of our approach.

2 Related Work

A related research domain is the propagating schemes, which mainly focuses on propagating the term frequency values or (given a query) the relevance score itself to find relevant pages in the web. For instance, given a query [17] propagates the relevance score between web pages connected with hyperlinks. On the other hand, the term frequency values can be propagated between neighboring pages [18]. *Qin et al.* [13] proposes a generic relevance propagation framework, which brings together techniques from [17] and [18]. The approach we present in this paper is also a term-frequency-propagation technique; however, unlike the above methods, the techniques we introduce

leverage the context provided by entries in an implicit quotation structure.

There has been much interest in tagging blog entries. [3] shows how to arrange tags into a hierarchy using clustering algorithms. A collaborative filtering method is used to automatically tag blog entries. In [10], given a blog entry, authors identify similar blog entries, collect tags assigned to these similar entries, and suggest these collected tags to user. CDIP is similar with [10] in that tags are borrowed from blog entries. However, unlike [10], our approach does not assume there exist manually assigned tags in similar blog entries. Furthermore, instead of blog-level similarity, CDIP identifies overlap (quotations) to extract logical links between blog entries and propagates tags along these links.

In the past few years, there has been growing research on blog analysis. Existing works leverage either structure information of blogosphere or contents of blog entries. [19] combines the structural information of blogosphere with content information to generate blog ranking given user queries. [8] proposes techniques to discover spatiotemporal theme patterns from blogs. [12] develops an algorithm to discover the topic development patterns and these pattern are used to segment blog archives. [11] studies to identify hot topics from blogs by identifying the role (such as agitator and summarizer) of bloggers in blog-based conversations. [2] propose a technique for inferring information propagation on blogs through embedded hyperlinks. These differ from our work in two major aspects. First, while existing methods rely on the explicit links to extract structural information [19, 2, 11], in our approach we aim to also leverage *quotations (or reuses)* among blog entries. Secondly, content-based analysis generally leverages keywords of blogs to compute similarity between them [12] or identify important keywords [8]. In our approach, instead of being limited to single entires, blog entries are enriched with keywords from contextually related entries and these enriched keywords are used for indexing and/or mining.

3 Representation of Blog Entries

In this paper, without loss of generality, we represent each blog entry using a keyword/tag-vector. For each entry, b_i, we construct a vector of weights, $P_i = \{w_{i,1}, w_{i,2}, \ldots, w_{i,n}\}$, where n is the number of keywords/tags in all entries concerned. Naturally, as conventional, stop words are removed and a stemmer is employed.

The weight of the entries in a keyword-vector can be term frequencies in the blog or TF-IDF weights capturing not only term frequencies, but also distribution of the keywords in the blogosphere. We note that keyword weights can incorporate additional domain knowledge (such as domain specific ontologies or namespaces) whenever applicable. Given these keyword vectors, it could also be possible to first mine for independent topics in the collection using a content-only scheme (such as *latent semantic indexing (LSI)*-based method [4]). However, the dynamic nature of the blogosphere causes these statistically-based techniques to miss important (but statistically insignificant relative to the entire blogosphere) developments. Nevertheless, if a representative corpus is available (e.g., technology), keyword weights can reflect importance of the words across related (in time or by reference) entries in different blogs. In this paper, we do not explore this option further, but leave it as future work.

4 Quotation and Re-use Detection

As shown in Figure 1, CDIP relies on quotation detection for establishing logical relationships between blog entries. Since applying quotation detection and content enrichment would be costly when applied across the entire blogosphere, we assume that an initial filtering step is applied to identify a candidate set of blog entries to be displayed to the user.

Reuse detection is an important problem that has implications in various application domains, including copy (*plagiarism*) detection and biological sequence mining. A particular challenge in reuse detection is that re-use can happen at different levels and detecting different types of reuses can require different techniques. In RECAP [9], Metzler *et al.* focus on the problem of *information flow* analysis and target identification of finer-granularity distinction between different reuse patterns: (a) identical modulo formatting, (b) sufficient overlap (common source), (c) some, common knowledge, overlap, and (d) no overlap. In particular, [9] showed that the word overlap measure was competitive when identifying reuse among sentences as opposed to general topic similarity at the all-of-document level for topical similarity and fact reuse.

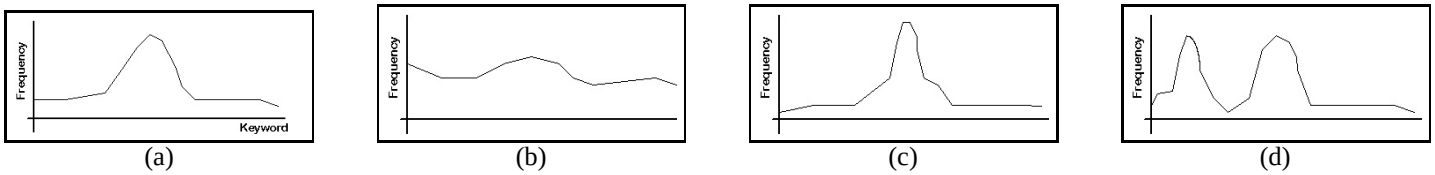
In word overlap, given a query sentence Q and a candidate sentence R, the level of similarity is defined as

$$S(Q,R) = \frac{|Q \cap R|}{|Q|}.$$

Metzler *et al.* also experimented with an IDF word overlap measure where keyword weight is readjusted by inversed document frequency. An IDF word overlap measure is given as

$$S(Q,R) = \frac{|Q \cap R|}{|Q|}\Big(\sum_{w \in Q \cap R} log\frac{N}{df_w}\Big).$$

Given such sentence-level evidence, RECAP [9] suggests building document scores by combining individual sentence-to-sentence scores in a bottom-up manner, thus better distinguish document pairs that share facts from those that simply have a topical similarity.



(a) (b) (c) (d)

Figure 2. (a) The original frequency distribution of a blog entry; (b) propagation process rendered the blog entry less discriminatory; (c) propagation process rendered the blog entry more focused; (d) propagation rendered the blog entry more general (but, still discriminatory)

Although this scheme is shown to be promising, a particular challenge, when applied in real-time (during the query-processing stage) is that sentence-level matching and evidence collection can be costly. Therefore, in [7], we proposed an efficient and incremental algorithm for scalable quotation detection across large blog and news collections.

5 Sharing vs. Individuality

Let us consider a set $B = \{b_0, \ldots, b_m\}$ of blog entries that are returned during the initial filtering phase. Since all of these entries are answers to the same query, naturally, they are expected to be related, especially regarding the keywords that were part of the query. In other words, for each blog entry, $b_i \in B$, the corresponding vector of keyword or tag weights, $P_i = \{w_{i,1}, w_{i,2}, \ldots, w_{i,n}\}$, are biased (i.e., high) for the query keywords. Naturally, these entries are differentiated from each other only through the weights of the non-query keywords and tags. Thus, any non-query keyword propagation across these blog entries will increase the risk of rendering these document too similar.

5.1 Frequency Distribution

A particular challenge in keyword and tag (weight) propagation is to ensure that, while enriching the documents with contextually relevant keywords, the discrimination power of the individual blog entries have not been significantly compromised. Figure 2, illustrates possible consequences of content enrichment (keyword propagation). Figure 2(a) shows the keyword distribution of a hypothetical blog entry. Figures 2(b) through (d) shows possible keyword distributions after the propagation. In Figure 2(b), the keyword propagation rendered the distribution of the keywords *flatter* (i.e., more uniform); i.e., the keywords in the document are rendered less discriminating. In (c), the distribution of the keywords changed in such a way that those keywords which were highly frequent became even more frequent; in other words, the entry became more *focused*. In (d), on the other hand, while the document became less focused around its original keywords, the distribution is still far away from flat and, hence, the updated entry is more *general* but *discriminating*. Note that it is essential that keyword propagation avoids cases where, the histogram of the blog entry is rendered flat as in Figure 2(b). Figure 2 illustrates the primary challenge in content enrichment process: while we would like to exchange contextually relevant keyword entries between blog entries that are related, we do not want these two entries become indistinguishable after the process. In other words, we would like to preserve the *difference* between these entries, although we are rendering them *similar* due to the increased sharing of keywords.

In this paper, we observe that understanding how specifically two given blog entries relate to each other can go a long way toward solving this challenge. In particular, a precise understanding of the relationship between the two entries can prevent *over-exchange* of keywords between them:

- exchanging keywords and tags among only those blog entries where there is explicit reuse (as opposed to performing exchange among all seemingly related entries) can ensure that content-enrichment follows the path of the information-flow,
- a quotation can provide the context in which two blog entries are related to each other, thus differentiating between related and unrelated keywords in the blogs, and
- understanding which blog entry is originally more general and which one is more focused can help preserving the relative shapes of the histograms after the propagation, thus preventing them from being rendered flat.

6 Quantifying the Degree of Focus

Identifying semantic similarity/dissimilarity between entries in a semantic hierarchy is a well studied problem [15, 14]. Most existing techniques extract these relationships either (a) from the information content of the terms in a given ontology computed over a large corpus [15] or (b) from the structure in a given hierarchy itself [14]. While these techniques have found application in various application domains where the semantic hierarchies are given, we note that there is a third and complementary piece of information (i.e., the textual content of the blog entries themselves) which can be exploited to extract the relationships, when such a semantic hierarchy is not known in advance.

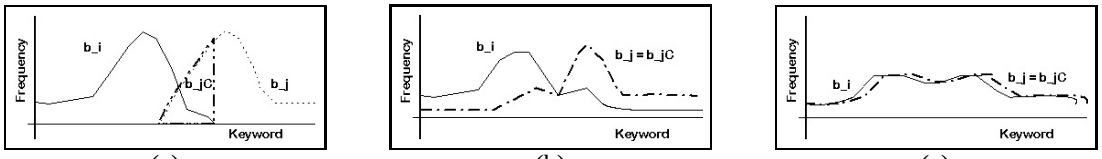


Figure 3. (a) The original keyword frequency distributions of blog entries b_i, b_j, and b_{jC}; (b) the keyword propagation rendered the blog entries similar, but distinct; (c) keyword propagation rendered the two blog entries indistinguishable

In our prior work, we presented content-driven techniques for measuring specialization/generalization degrees between entries in hierarchical structures, such as IS-A concept hierarchies [6] and message hierarchies, such as discussion board [5]. In this paper, we highlight that, a particular advantage of these content-driven techniques is that they do not assume any advance knowledge about the structure and, thus, can be used when only pairwise content (and reuse such as quotations) are available.

Given a pair of blog entries, b_i and b_j, one way to think of relative content of these two data entries is in terms of constraints imposed on them by their keyword compositions: the statement that "*entry b_i is more general than b_j*" can be interpreted as b_i being less constrained than b_j by its keywords. In a sense, in measuring degrees of generality/focus, we need to be able to *quantify* how well a given blog entry can be interpreted as the *disjunction* of the corresponding keywords. Relying on extended boolean model [16] (which aims to measure the degree of match between a given text and a query consisting of *or* and *and* logical connectives), we measured the degree of generality as follows:

$$G_{ij} = \frac{|\vec{b_i}|}{|\vec{b_{jC}}|},$$

where for any $\vec{x}$, $|\vec{x}|$ denotes the distance from the origin $\vec{\mathcal{O}}$, and b_{jC} is the common parts of b_i and b_j (See [6] and [5] for more detail).

Keyword or tag exchange between blogs, b_i and b_j, will most likely increase b_{jC} (i.e., the common base of the two blog entries). Thus, ensuring that the keyword distribution does not get flat as in Figure 2(b) requires that both G_{ij} and G_{ji} are preserved despite the increases in b_{iC} and b_{jC} due to their enrichment. Figure 3 visually depicts this process. In Figure 3(b), the keyword propagation is performed in a manner which preserves G_{ij}, thus (although keywords are exchanged) the two documents are still distinguishable, while in Figure 3(c), since, after the keyword propagation, $b_i \sim b_j (= b_{jC})$, G_{ij} has not been preserved, and consequently, keyword exchange rendered the two documents indistinguishable and largely indiscriminating. Thus, preservation of G_{ij} provides an upperbound for the possible content exchange between the blog entries.

In the following section we will show how to leverage and extend this for quotation-aware keyword propagation. Then, in Section 8, we will show that, while such focus preserving content-exchange between blog entries, b_i and b_j, enriches both blog entries with new content, it does not do so in the expense of flattening both documents.

7 Collection-Driven, yet Individuality-Preserving Tagging of Blog Entries

In this section, we first discuss how to propagate keywords, tags, and their weights, between blog entries in a given set B. In [6], we described the propagation in the form of a matrix.

7.1 Basic Propagation Matrix

Given a set of blog entries B, its corresponding propagation matrix, $\mathtt{M}_B$, is a $|B| \times |B|$ matrix, such that

- $\mathtt{M}_B$ is symmetric,
- diagonal values are of $\mathtt{M}_B$ all 0, and
- $\forall i, j \quad 0 \leq \mathtt{M}_B[i,j] = \alpha_{i,j}$.

Here, $\alpha_{i,j} = \alpha_{j,i}$ denotes the amount of keyword propagation between blog entries b_i and b_j.

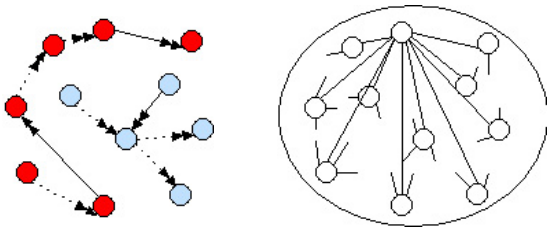
Given a propagation matrix, $\mathtt{M}_B$ and a "keyword"-"blog entry" matrix $\mathtt{T}_B$, the matrix $\mathtt{P}_B = \mathtt{M}_B\mathtt{T}_B$, describes the amount of keywords propagated between pairs of blog entries in B. Thus, the new *enriched* "keyword"-"blog entry' matrix is equal to

$$\mathtt{T}'_B = \mathtt{T}_B + \mathtt{P}_B = \mathtt{T}_B + \mathtt{M}_B\mathtt{T}_B = (\mathtt{I}_B + \mathtt{M}_B)\mathtt{T}_B$$

where $\mathtt{I}$ is the identity matrix. Note that since all diagonal values in $\mathtt{M}_B$ are zero, $\mathtt{I}+\mathtt{M}_B$ is such that all diagonal values are 1 and all non-diagonal entries are those in $\mathtt{M}_B$.

7.2 Focus Preserving Propagation

As highlighted in Section 6, such a propagation of keywords should change the composition of the keyword vectors of the entries ($\mathtt{T}'_B \neq \mathtt{T}_B$), yet *after the keyword propagation the keyword distributions of individual entries should not end up being be flattened due to the shared keywords*. In other words, the propagation matrix $\mathtt{M_B}$, should be such



(a) (b)

Figure 4. (a) Quotation-based propagation vs. (b) all-to-all propagation

that the relative generality of the pair of entries should be preserved, despite of the fact that there is larger keyword sharing between the blog entries. Since after the keyword propagation between entries, b_i and b_j, the two entries are located in a common keyword space, the propagation matrix $\mathtt{M_B}$, should ensure the modified equality

$$G_{ij} = \frac{|\vec{b_i}|}{|\vec{b_{j_C}}|} = \frac{|\vec{b'_i}|}{|\vec{b'_j}|} = \frac{|\vec{b_i} + \alpha\vec{b_j}|}{|\vec{b_j} + \alpha\vec{b_i}|} = G'_{ij},$$

where $\alpha = \mathtt{M}_B[i,j] = \mathtt{M}_B[j,i]$ is the propagation degree between the blog entries, Note that it is possible to compute the appropriate propagation degree by solving this equation for α. Note also that, when the keywords of b_i and b_j are identical, if we try to solve for $G_{i,j} = G'_{ij}$, we get $\alpha = 0$ as a solution. Consequently, keyword propagation is not applicable when two entries, b_i and b_j, have the same set of keywords. This provides a stopping condition for keyword propagation process; given a set B of blog entries, focus-preserving propagation needs to be applied at most $|B|$ times.

8 Experimental Evaluation

We have evaluated CDIP approach developed in this paper through experiments. In terms of efficiency, we measured the gains in propagation time due to the use of *quotations* as opposed to *all-to-all* propagations (Figure 4). On the other hands, in terms of effectiveness, we showed that quotation-based propagation helps preserving the individuality of the blog entries, and explicit preservation of the focus through *semantic-focus analysis* helps improve the preservation of distinctness of blog entries.

8.1 Tag Enrichment Analysis

Setup: For the experiments presented in this section, we used 300 entries that consist of 30 topics (that is, 10 entries per each topic) and were collected from Google Blogsearch and Google News on May 14, 2007. By using the sentence-level based quotation detection discussed in Section 4, we

	KL distance between pairs of entries belonging to the same topic	KL distance from uniform distribution
Initial	7.754	7.378
quotation+preserve	5.494	7.074
quotation+const (0.1)	5.483	7.062
quotation+const (0.5)	5.475	7.060
alltoall+preserve	0.001	6.365
alltoall+const (0.1)	0.001	6.347
alltoall+const (0.5)	0.001	6.346

Table 1. The average KL distance between uniform distribution and entries and between pairs of entries.

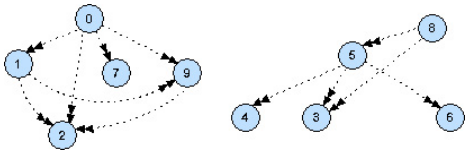


Figure 5. The quotation structure of the 10 entries used in these experiments (the entries are results to an initial keyword query: "tivo"). The edges denote the quotations identified using the algorithm presented in Section 4. The arrows denote the focus direction using the algorithm presented in Section 6.

obtained the quotation structure of the 300 entries that contains 481 edges between blog entries. On the other hands, there were 1350 edges in all-to-all propagation structure.

Experimental Results: We first measure the average KL-distance between pairs of entries belonging to the same topic. As can be seen in the Table 1, the distance between pairs of blog entries was reduced after keyword propagation compared with initial case. However, quotation driven propagations maintained certain degree of distinction between the entries after propagation. Especially, it is observed that the KL-distance between pairs of blog entries significantly drops when all-to-all propagation used, which indicates that all-to-all propagations almost lost the individuality of entries and is not appropriate for tagging and searching of blog entries. Then, we computed the average KL-distance of resulting keyword distribution after propagation from a uniform distribution; thus, a high value means more discriminating keyword composition. As shown in the the Table 1, the highest distance is obtained when quotation-driven propagation is applied. Furthermore, it is observed that although the change in the constant used for focus-unaware propagation is different in the Table 1, there is hardly any difference between the two result sets. Thus, as long as focus preservation is not employed, the constant used for focus-unaware propagation has little impact on the propagation result. These results on two experiments imply that the proposed approach in this paper can preserve the original entries' discriminating power after context enrichment.

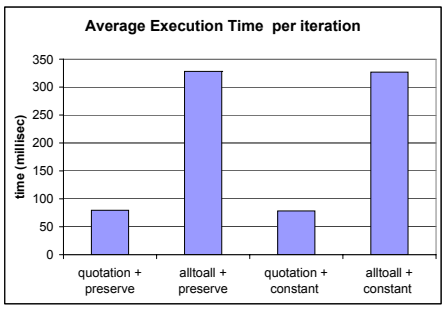


Figure 6. Per iteration execution time for quotation-based and all-to-all propagation schemes. Note that, in terms of execution time, there is no discernible difference between whether a focus-preserving scheme or a constant-propagation scheme is used.

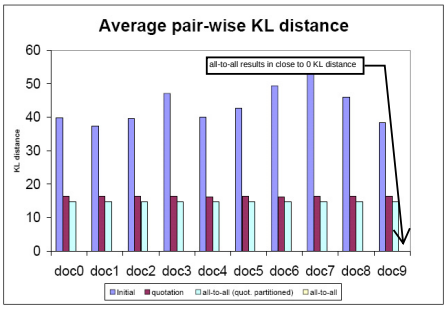
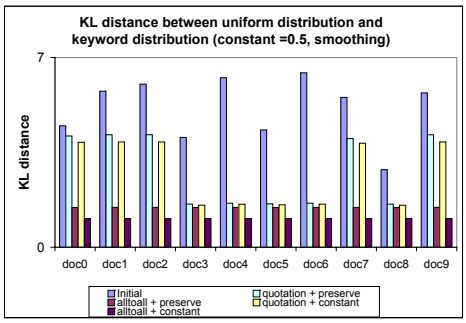
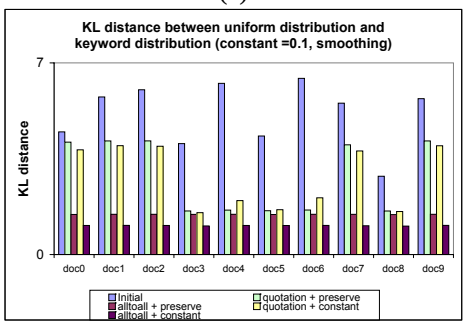


Figure 7. Content enrichment: the difference between keyword distributions



(a)



(b)

Figure 8. Distance from uniform distribution: Closer "0" means more uniformly distributed; i.e., keywords are less discriminating. Figure (a) corresponds to constant propagation weight 0.5, while Figure (b) corresponds to 0.1. As can be seen here the use of different constants do not improve the result

8.2 Case Study: A Detailed Look at the Quotation Analysis

Setup: For the experiments presented in this section, we used the query keyword "tivo" to collect entries from *Google News* and *Google Blogsearch* and selected randomly 10 of the returned documents. Figure 5 shows the quotation structure of these 10 entries obtained using the sentence-level evidence based quotation detection discussed in Section 4. Note that although all of these documents were results for the same query, the quotation structure is relatively sparse and splits the 10 entries immediately into two separate reachability partitions. The figure also shows the results of the *semantic-focus* analysis as presented in Section 6. Note that, as shown in the figure, the *semantic-focus* analysis process resulted in tree-like, acyclic relationships between the entries. In the left-most partition, the entry "0" seems to be the most general, while "2" seems to be the most focused. In the rest of the section, we will evaluate the use of these quotations and the focus-preservation on keyword and tag propagation on these data sets.

Experimental Results: First, we compare the execution times of the *quotation*-based and *all-to-all* propagation schemes. As shown in Figure 6, each iteration of the content propagation algorithm takes significantly less time when only quotations are considered.

Figure 7 shows the content enrichment results after an entire enrichment cycle. The figure measures the average difference between input keyword distributions using the KL-divergence metric. As can be seen in the figure, for all documents, after keyword propagation, the distance from uniform was reduced. When the keyword-propagation was performed all-to-all, the difference between the blog entries was lost. Performing quotation-driven propagations, however, maintained certain degree of distinction between the entries. Performing all-to-all (but within each quotation-partition separately) preserved distinction between entries. Following the quotations (instead of performing all-to-all in each partition) was most effective in maintaining distinctions between entries.

Since there is a non-negligible drop in distinctness of the documents according to Figure 7, we need to look at more carefully to the resulting distributions. Figure 8 provides a detailed view of the keyword distributions for the 10 entries. Especially, these two plots show the distance of the resulting keyword distribution from a uniform distribution; thus, a high value means more discriminating keyword composi-

Entry "2"		Entry "3"	
Top-5 Original Tags	Top-5 Borrowed Tags	Top-5 Original Tags	Top-5 Borrowed Tags
Tivo	Feature	Tivo	Subscribe
Video	Download	Video	Forward
Service	Year	Content	Statement
Subscribe	Corporation	Television	Offer
CBS	Quote	CBS	Access

Table 2. Tag propagation samples

tion. The observations can be summarized as follows:

- When we look at Figure 8(a), we see that there are two classes of results. For entries,"0", "1", "2", "7", and, "9" (which are in a quotation group together according to Figure 5), quotation-based propagation results in a large degree of distance from the uniform. On the other hand, for the second partition of entries (i.e., "3", "4", "5", "6", and "8") the result is much closer to uniform, even though the entries are initially more discriminating. Nevertheless, it is important to note that, for both partitions, quotation-based propagation schemes maintain a higher degrees of discrimination power than all-to-all propagation.

- The explicit focus preservation helps in preventing the final distribution to resemble uniform distribution. This is especially visible for all-to-all propagation, where the risk of resulting in a uniform-distribution is higher. Therefore, we can conclude that, when there are no quotations to leverage to identify information flow patterns between the entries, focus preservation may help maintaining distinctness of the entries.

Finally, Table 2 provides sample propagation results for blog entries, "2" and "3", for the reader's quick reference.

9 Conclusion

In this paper, we highlighted that frequent content overlaps between blog entries are both (a) significant challenges against effective retrieval and (b) yet, if leveraged effectively, can help alleviate part of the problem by enabling contextually informed tag sharing among related blog entries. Based on this observation, we presented *reuse-detection*, *content analysis*, and *keyword and tag propagation* schemes which let blog entries share significant keywords and tags, while preserving their individual focus for discriminating searches. Based on these, we developed CDIP which enables collection-driven automated tagging and evaluated the proposed method for effectiveness.

References

[1] David Sifry's Blog, http://www.sifry.com/alerts/.

[2] E. Adar and L.A. Adamic. Tracking Information Epidemics in Blogspace. Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence, 2005.

[3] C. H. Brooks and N. Montanez. Improved Annotation of the Blogosphere via Autotagging and Hierarchical Clustering. In *WWW*, 2006.

[4] T. Hofmann. Probabilistic latent semantic analysis. In *UAI'99*, Stockholm, 1999.

[5] J.W. Kim, K.S. Candan, and Mehmet E. Dnderler Topic segmentation of message hierarchies for indexing and navigation support. In *WWW*, 2005.

[6] J.W. Kim and K.S. Candan. CP/CV: Concept Similarity Mining without Frequency Information from Domain Describing Taxonomies. In *CIKM*, 2006.

[7] J.W. Kim, K.S. Candan, and J. Tatemura. Identification and Organization of Inter-related Blog and News Entries for Effective Exploration. submitted, 2007.

[8] Q. Mei, C. Liu, H. Su, and C. Zhai. A Probabilistic Approach to Spatiotemporal Theme Pattern Mining on Weblogs. In *WWW*, 2006.

[9] D. Metzler, Y. Bernstein, W.B. Croft, A. Moffat, and J. Zobel. Similarity Measures for Tracking Information Flow. In *CIKM*, 2005.

[10] G. Mishne. AutoTag: A Collaborative Approach to Automated Tag Assignment for Weblog Posts. In *WWW*, 2006.

[11] S. Nakajima, J. Tatemura, Y. Hino, Y. Hara, and K. Tanaka. Discovering Important Bloggers Based on a Blog Thread Analysis. *Workshop on the Weblogging Ecosystem*, 2005.

[12] Y. Qi and K.S. Candan. CUTS: CUrvature-based development pattern analysis and segmentation for blogs and other Text Streams. In *HYPERTEXT*, 2006.

[13] T. Qin, T. Liu, X. Zhang, Z. Chen, and W. Ma. A Study of Relevance Propagation for Web Search. In *SIGIR*, 2005.

[14] R. Rada, Mili, H., Bicknell, E., and Blettner, M. Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man and Cybernetics, 19(1), 1989.

[15] P. Resnik. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. JAIR, Vol.11, 1999.

[16] G. Salton, E.A. Fox, and H. Wu. Extended Boolean information retrieval. CACM, 26(11). 1983.

[17] A. Shakery and C. Zhai. Relevance Propagation for Topic Distillation UIUC TREC-2003 Web Track Experiments. In *TREC*, 2003.

[18] R. Song, J.R. Wen, S. Shi, G. Xin, T.Y. Liu, T. Qin, X. Zheng, J. Zhang, G. Xue, and W.Y. Ma. Microsoft Research Asia at Web Track and Terabyte Track of TREC 2004. In *TREC*, 2004.

[19] B. Tseng, J. Tatemura, and Y. Wu. Tomographic clustering to visualize blog communities as mountain views. In *WWW'04 Workshop on the Weblogging Ecosystem*, 2005.