# Topic Exploration and Distillation for Web Search by a Generalized Similarity Analysis

Xiaoyu Wang, Hongwei Wu, Li Wei, Aoying Zhou

Department of Computer Science and Engineering
Fudan University, Shanghai, 200433, P. R. China
{xiaoyuwang | hweiwu | lwei | ayzhou}@fudan.edu.cn

**Abstract**

Topic distillation is the process of finding representative pages relevant to a given query. Well-known topic distillation approaches such as the HITS algorithm have shown to be useful in identifying high quality pages of the most popular topic within a query specific graph of hyperlinked documents. Many succeeding researchers focus on augmenting HITS with further content analysis to alleviate the steady deterioration of distillation quality suffered by HITS. In this paper, we attempt to revisit the behavior of HITS from a different point of view. Namely, a similarity-based analysis model is applied to observing the distillation procedure. By defining a generalized similarity, an algorithm is proposed, which can improve the quality of distillation only using the information of hyperlinks. A topic exploration function is also integrated in the algorithm framework, which enables end-users to search less popular topics when multi-topics are involved in queries. The experimental results reveal two benefits from the new algorithm: the improvement of distillation quality without utilizing any content information of pages, and an additional ability to explore the topics emerging in the query results.

**Keywords:** link analysis, web search, topic distillation, and topic exploration.

# 1 Introduction

A number of papers [BP98, Klei98, CDG98 and BH98] have considered the use of hyperlinks for the purpose of Web search in recent years. In particular, these papers consider the extent to which hyperlinks between World Wide Web documents can be used to determine the relative authority values of these documents for various search queries. This process of finding quality pages is called *topic distillation* [BH98].

The well-known HITS algorithm was proved to be useful for topic distillation by a connectivity analysis in hyperlink environments [Klei98]. While this algorithm works well for some queries, continual experimental results reported by several researchers [CDG98 BH98] reveal a steady deterioration of distillation quality (*topic drift* problem) through the last few years. They extended the HITS algorithm to avoid topic drift by employing the textual information of documents. Other works of link analysis such as Google combining link-based ranking with page text and anchor text in undisclosed ways, and keeps tweaking the combination.

Many had thought that those text-linkage hybrid approaches would put an end to the pure link-analysis, but the situation is clearly more like an ongoing arms-race. In fact, an extra burden of fetching the entire pages as well as heavy tasks of textual processing is required in those hybrid approaches. Dropping pages that are judged to be "outliers" using threshold of textual similarity improves precision, but may reduce the recall if some of the pages with smaller similarity are pruned. This may not be a problem for broad queries but could be serious for narrower ones. Also, none of those approaches provided a solution to the queries involving multi-topics.

Linkage information between documents, as ample clues for topic distillation and exploration, could be more harvested than we had thought. In this paper, we initiate an intensive study of the HITS algorithm with a similarity-based analysis model, and then propose a new topic distillation algorithm. It attempts to resolve the problem of *topic drift* employing link-analysis without utilizing any further textual information in pages. The proposed algorithm also enables end-users to discover all the possible topics (topic exploration) occurring in queries. We compare the performance of our approach with the HITS algorithm on various query topics, and the empirical evaluation shows that our algorithm is effective.

The paper structured as follows. Section 2 shows the connection with the related work. Section 3 describes the HITS algorithm and studies the behavior of HITS with a similarity-based analysis model. Section 4 proposes the new similarity definition that will be used in our algorithm. Section 5 gives the framework of our algorithm. In section 6, we present some experimental results of our algorithm as well as HITS and evaluate them. Section 7 then summarizes our work and outlines the future work.

# 2. Related Work

The analysis of link structure with the goal of understanding informational organization has been an issue in a number of overlapping research field. In this section, we review some of the previous approaches and point out the connections between our work and them.

Bibliometrics is the study of written documents and their citation structure. Research in bibliometrics has long been concerned with the use of citation to identify core sets or clusters of articles, authors, or journals of particular fields of study [WM89, Sm73]. Small and Griffith developed co-citation analysis [SG74] as a method for measuring the common intellectual interest between a pair of documents. Many researchers have applied the

co-citation analysis to the exploration of the implicit semantic structure of the WWW. Larson [Lar96] discussed on applying bibliometrics to the World Wide Web in 1996. Since then, a series of applications of bibliometrics were made to the hyperlinked environments, with the purpose of identifying the most authoritative pages related to the user query.

One of the most well known works might be the HITS algorithm proposed in [Klei98]. Several researchers have extended this basic algorithm. Chakrabarti et al [CDG98] weighted links based on the similarity of the text surrounded the hyperlink in the source document to the query that defined the topic. Bharat and Henzinger [BH98] made several important extensions. First, they weighted the documents based on their similarity to the query topic. Second, they counted only links between documents from different hosts, and average the contribution of links from any given host to a specific document. Chakrabarti [Cha01] proposed a uniform fine-grained model for the Web in which pages are represented by their tag trees so that mutual endorsement between hubs and authorities involve the segments of DOM tree.

The common effort that those extended approaches made is to solve the problem of *topic drift* encountered using HITS [Klei98]. While different techniques are employed in those extended works, they all utilized the content in the documents. The main goal of our approach is the same as those extended works. However, our algorithm only employs the hyperlink structure information to solve the *topic drift* problem, thereby avoiding an extra burden of page fetching and content analyzing.

Not only topic distillation approaches have intensively analyzed the hyperlinks of web pages, but many others have used inter-document linkage to compute useful data on the Web as well. Botafago [Bot93] proposed a graph-based algorithm for clustering hypertext that uses link information; he proposed the number of independent paths between nodes as a measure of similarity. Pirolli [PPR96] have combined both the link topology and textual similarity between items as well as usage data collected by servers and page meta-information like title and size. Chen [Che97] has proposed generalized similarity analysis that combines hypertext linkage, content similarity, and browsing patterns or usages. Chen and Czerwinski [CC98] have exploited generalized similarity analysis along with latent semantic indexing and pathfinder network scaling to develop an integrated framework for spatial organization of information, browsing and searching.

While those works attempt to integrate the linkage of documents into the similarity definition, they are applied to clustering purpose. The similarity definition in our work is oriented in a different direction, namely, to use similarity as a mean of analyzing the behavior of topic distillation process. Furthermore, unlike those previous definitions of pair-wise similarity, the proposed similarity definition in our work is a more generalized one that captures the relationships in sets with arbitrary cardinalities.

Those papers concerned with clustering are some of the works in IR on supporting topic exploration. The goal of topic exploration is to locate a set of documents dealing with the user's topic interest. Previous topic distillation approaches based on HITS assume such a set and find quality documents within it. Such a limitation of HITS-like algorithms makes them difficult to explore all possible topics when processing ambiguous queries.

Davison et al [DGK00] have built a prototype called Disco Web that can discover more topics than HITS. By looking at a larger set of eigenvectors, Davison claimed to find clusters of web pages that are more interesting than those extracted by the principal eigenvector, and used heuristics to extract a globe ranking as well as local

rankings given by each eigenvector. Though the detailed techniques are not presented in [DGK00], the effort of computing a number of eigenvectors is extremely hard. Also, the heuristics of finding possible topics from the set of eigenvectors must be tedious. Unlike Davison et al, we integrate a topic exploration function in the algorithm, which enables the returned results to be ranked according to each possible topic of the query only with the principal eigenvector of each topic sub-graph.

# 3 Intensive Study of HITS Algorithm

This section reviews the HITS algorithm and discusses the behavior of it with a similarity-based analysis model, thereby indicating the embarrassments it suffers.

## 3.1 Review of HITS Algorithm

In the algorithm, two scores for each document have to be calculated: a hub score and an authority score. A document which points to many others is a good hub, and a document that many documents point to is a good authority. Transitively, a document that points to many good authorities is an even better hub, and similarly a document pointed to by many good hubs is an even better authority.

Starting from a user-supplied query, the algorithm assembles a root set $R$ of pages returned by a search engine on that query. The root set $R$ is then expanded into a larger base set $T$ by including pages that point to any page in it, and pages that are pointed to by any page in it. Links which are between pages in the base set with the same domain play a role of navigation and should not be considered.

Suppose the resulting graph is $G = (V, E)$. Each page $p$ in $V$ has a pair of non-negative weights $<a_p, h_p>$ where $a_p$ is the authority score and $h_p$ is the hub score. Before the start of the algorithm, all the $a$- and $h$-values are initially set to 1. the value of $a_p$ is updated to be the sum of $h_q$ over all pages $q$ that link to $p$: $a_p = \Sigma_{q|q \to p} h_q$. In a strictly dual fashion, for a page $p$, its hub weight is updated via $h_p = \Sigma_{q|p \to q} a_q$. The algorithm repeats those steps a number of times, at the end of which it generates the rankings of the pages by their hub and authority scores.

If the graph is represented with $A$ in the adjacency matrix format (i.e., matrix $A$ whose entry of $(i, j)$ is equal to 1 if and only if $i \to j$, and is 0 otherwise), then the above operation can be written simply as $a \leftarrow A^T h \leftarrow A^T A\ a = (A^T A)\ a$, and $h \leftarrow Aa \leftarrow AA^T h = (AA^T)\ h$, interspersed with scaling to set $| a | = | h | = 1$. Thus the vector $a$ after multiple iterations is precisely the result of applying power iteration to $A^T A$, and a standard result in linear algebra [GL89] tells that this sequence of iterates, when normalized, converges to the principal eigenvector of $A^T A$. Similarly, the sequence of values for the normalized vector $h$ converges to the principal eigenvector of $AA^T$.

## 3.2 Analyzing HITS Algorithm Using Similarity-based Model

The behavior of the HITS algorithm is described with a similarity-based analysis model in this sub-section. The underlying idea is that pages which co-cite (share common out-links) or are coupled (share common in-links) are with high probability to be related in semantic. Let us number the pages in the base set $T$ as $\{1, 2 \ldots n\}$, and represent each page in $T$ as two $n$-dimension vectors: $v^{out}$ and $v^{in}$. Given a page $p$ in $T$, the $i^{th}$ component of vector $v_p^{out}$ is 1 if and only if the page $p$ points to page $i$ and is 0 otherwise. Identically, the $i^{th}$ component of vector $v_p^{in}$ is 1 if and only if the page $p$ is pointed to by page $i$ and is 0 otherwise. It is assumed that there is no linkage from page $p$ to itself (The $i^{th}$ component of vector $v_i$ is 0). Inner product is adopted to measure the semantic relationship

between a pair of pages. In particular, $Similarity^{in}(p, q) = v_p^{in} \bullet v_q^{in}$, and, $Similarity^{out}(p, q) = v_p^{out} \bullet v_q^{out}$.

Obviously, the $Similarity^{out}(p, q)$ captures the common out-links shared by $p$ and $q$ whereas $Similarity^{in}(p, q)$ captures the common in-links shared by $p$ and $q$ when $p$ is not equal to $q$. The notation of *Similarity* in this context could correspond to the *co-citation strength* in co-citation analysis. When $p$ is equal to $q$, the similarity value indicates the total in-link number or out-link number of that page. Actually, the similarity matrix whose entry $(i, j)$ is the $Similarity(i, j)$ could just be seen as the basis of power iteration in the HITS algorithm. More formally, It could be stated in the theorem below.

**Theorem 3.2.1** There exist similarity matrices $S^{in}$ and $S^{out}$ whose entry $S^{in}(i, j)$ and $S^{out}(i, j)$ are equal to $Similarity^{in}(i, j)$ and $Similarity^{out}(i, j)$ respectively, such that, $S^{in}$ is equal to $A^TA$, and $S^{out}$ is equal to $AA^T$, where $A$ is the adjacency matrix of the given query graph in the HITS algorithm. ■

By the definition of the $(S^{in})^k$ and $(S^{out})^k$ matrices, $k$-order similarities $Similarity_k^{out}(i, j)$ and $Similarity_k^{in}(i, j)$ can be defined, where $Similarity_k^{in}(i, j) = (S^{in})^k (i, j)$ and $Similarity_k^{out}(i, j) = (S^{out})^k (i, j)$. If $k = 1$, $Similarity_k^{in}(i, j)$ and $Similarity_k^{out}(i, j)$ fall into the basic definitions of the similarity $Similarity^{in}(i, j)$ and $Similarity^{out}(i, j)$ which will be denoted by $Similarity_1^{in}(i, j)$ and $Similarity_1^{out}(i, j)$ respectively in the rest of our paper.

After the $k^{th}$ iteration of the HITS algorithm, the authority vector $a_k$, and hub vector $h_k$ are the unit vectors in the direction of $(A^TA)^k u$ and $(AA^T)^k u$ respectively, where $u$ is ones vector $\{1, 1, 1 \ldots 1\}$. Following from Theorem 3.2.1, we have that $a_k(i) = \sum_{j=1}^{n} Similarity_k^{in}(i, j)$ and $h_k(i) = \sum_{j=1}^{n} Similarity_k^{out}(i, j)$.

Now, the undirected weighted graph $G_s^{in}$ can be defined as follows. The vertex set of the graph is the set of nodes in the base set $T$. An edge between two nodes $i$ and $j$ is drawn if the $Similarity_1^{in}(i, j)$ is not equal to 0. The weight of the edge is $Similarity_1^{in}(i, j)$. When $i = j$ and $Similarity_1^{in}(i, j)$ is not equal to 0, a circle edge is drawn at that node. We perform a traverse walk from node $i$ to node $j$ on graph $G_s^{in}$. If there exists $m$ distinct $k$-edge-paths between the two nodes $i$ and $j$, then a score value $\sigma_k^{ij}$ is assigned for each $k$-edge-path using the product of the all edge weights in that corresponding path. From the power theorem of adjacency matrix on graph theory, we have $Similarity_k^{in}(i, j) = \sum_{t=1}^{m} \sigma_k^{ij}(t)$. Graph $G_s^{out}$ and the $Similarity_k^{out}(i, j)$ can be defined in a similar way. The following lemma and theorem summarize the discussion above:

**Lemma 3.2.1** The $k$-order similarity of two pages $Similarity_k^{in}(i, j)$ is equal to $\sum_{t=1}^{u} \sigma_k^{ij}(t)$ if there exist $u$ $k$-edge-paths between node $i$ and $j$ in graph $G_s^{in}$. And the $k$-order similarity of two pages $Similarity_k^{out}(i, j)$ is equal to $\sum_{t=1}^{v} \sigma_k^{ij}(t)$ if there exist $v$ $k$-edge-paths between node $i$ and $j$ in graph $G_s^{out}$. ■

**Theorem 3.2.2** Given $n$-page base set, the $i^{th}$ component of authority vector $a_k$ after $k^{th}$ iteration of the HITS algorithm is equal to $\sum_{j=1}^{n} \sum_{t=1}^{u} \sigma_k^{ij}(t)$ where $u$ is the number of $k$-edge-paths between node $i$ and $j$ in the associated graph $G_s^{in}$. And the $i^{th}$ component of hub vector $h_k$ after $k^{th}$ iteration of the HITS algorithm is equal to $\sum_{j=1}^{n} \sum_{t=1}^{v} \sigma_k^{ij}(t)$ where $v$ is the number of $k$-edge-paths between node $i$ and $j$ in the associated graph $G_s^{out}$. ■

According to the theorems above, the essential reason of the embarrassments suffered by HITS is obvious.

The score $\sigma_k$ of $k$-edge-path from a node to another node in the weighted graph can be seen as $k$-radius transfer similarity of two pages in the base set. For many query-specific graphs with strong tendency to drift, such a transfer similarity always tends to be distorted, thereby further distorting the high-order similarity. (A simple example will be shown in next section.) Following Theorem 3.2.2, *Topic drift* problem occurs when a number of high order similarities between pair-wises are misled.

Theorem 3.2.2 also tells us that the HITS algorithm is an effort to find the pages with the highest $n$-order similarities to all the others in terms of the "global" or "overall" structure of the weighted graph when the similarity order is equal to the node number of the graph. Such an essence of the HITS algorithm makes it difficult to explore less popular topics involved in the given queries. We refer this problem as *topic missing*.

# 4 Inter-Page Similarities from Confidence of Association Rules

As analyzed in the previous section, the behavior of the HITS algorithm is based on the pair-wise relationships (similarities) of pages, which captures the number of pages that co-cite the pair, or the number of pages that the pair co-cite. In the following discussion, we will only concentrate on the co-cited similarity ($Similarity_1^{in}$), which leads to the computation of the authorities, and one can similarly understand the co-citing similarity ($Similarity_1^{out}$) when computing hubs. Taken over all pairs of pages, the co-citation similarity serves as a compact representation of the hyperlinked structure in query graph, thereby deriving the undirected weighted graph $G_s^{in}$. To illustrate the distortion of high order similarity of the HITS algorithm in an intuitive way, an example is given below.

**Example 4.1** Considering two distinct relationships among three nodes $i$, $j$ and $k$ (see Figure 1), obviously, the relationships between nodes $i$ and $j$ in part A should be less close than those in part B. The one-order similarities associated with those nodes in part A and part B are listed respectively in Table 1.

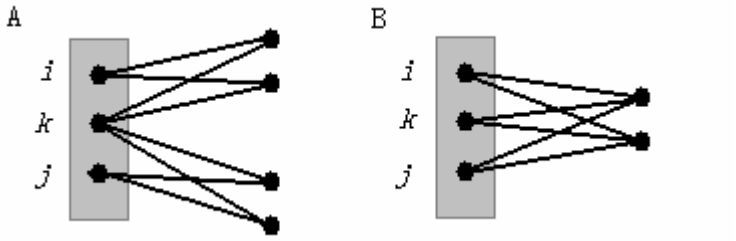


Figure 1. Two distinct relationships of nodes $i$, $j$ and $k$

**Table 1.** One-order similarities of the nodes in Figure 1

| Part | *A* | B |
|---|---|---|
| $Similarity_1^{in}$ $(i, i)$ | 2 | 2 |
| $Similarity_1^{in}$ $(i, k)$ | 2 | 2 |
| $Similarity_1^{in}$ $(k, k)$ | 4 | 2 |
| $Similarity_1^{in}$ $(k, j)$ | 2 | 2 |
| $Similarity_1^{in}$ $(j, j)$ | 2 | 2 |

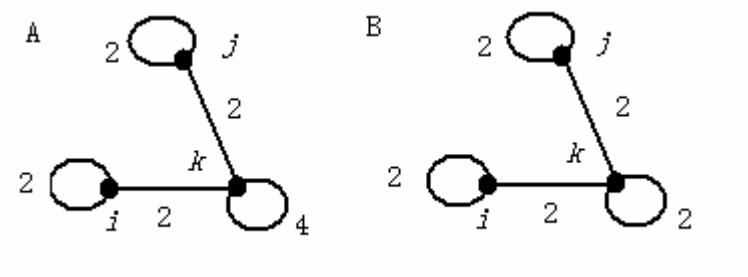


Figure 2. The path between nodes *i* and *j* in Figure 1

Following Lemma 3.2.1, the high order similarity of two nodes can be computed by the product of the weights of all edges (one-order similarities) in the path between them. In this example, the associated graph $G_s^{in}$ is shown in Figure 2 where the path between node *i* and *j* is illustrated in part A and B corresponding to the relationships of nodes in part A and B of Figure 1 respectively. To contain all edges (including self-loop edges) in the path, the similarity order is set to 5. The 5-order similarity of *i* and *j* is computed as the follows.

Part A: $Similarity_5^{in}(i, j)$

$= Similarity_1^{in}(i, i) \bullet Similarity_1^{in}(i, k) \bullet Similarity_1^{in}(k, k) \bullet Similarity_1^{in}(k, j) \bullet Similarity_1^{in}(j, j)$

$= 2 \times 2 \times 4 \times 2 \times 2 = 64.$

Part B: $Similarity_5^{in}(i, j)$

$= Similarity_1^{in}(i, i) \bullet Similarity_1^{in}(i, k) \bullet Similarity_1^{in}(k, k) \bullet Similarity_1^{in}(k, j) \bullet Similarity_1^{in}(j, j)$

$= 2 \times 2 \times 2 \times 2 \times 2 = 32.$

An opposite judgment of the relationship between *i* and *j* is derived from the 5-order similarities above. In particular, the 5-order similarity of node *i* and *j* in part *A* is twice that of node *i* and *j* in part *B*. ■

According to Theorem 3.2.2, the authorities of pages are essentially determined by the *n*-order similarities of all pair-wise nodes in *n*-node graph $G_s^{in}$. The example 4.1 is just a simple case indicating the distortion of high-order similarity caused by the unreasonable transfer similarities. Thereby, to distill more accurate results, the similarity definition of pair-wise nodes using co-citation strength in the HITS algorithm should be reconsidered.

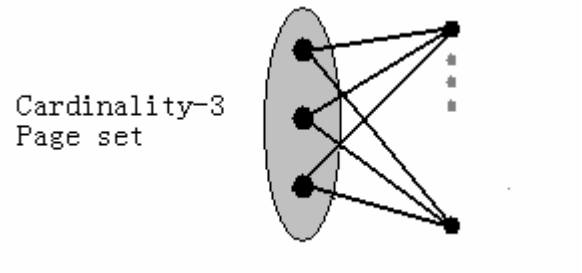


Figure 3. Co-citation strength based on arbitrary cardinality set

In order to alleviate the distortion of high-order similarities, we propose a generalization of the co-citation similarity in which sets of cardinality above two are considered, as shown in Figure 3.

The problem of finding a co-citation involved more than two nodes could be regarded as the generation of a frequent itemset in association rules mining. In our context, the set consisting of all pages concerned corresponds

to the set of items and a citation of each page in the base set corresponds to a transaction. A frequent itemset found using the association rule discovery algorithm corresponds to a set of pages that have more than one cited page in common. The frequent itemsets capture the relationships among items of size greater than or equal to 2. To measure the relationship among the pages in one frequent itemset, one possible way is to use the support of each frequent itemset which essentially is the co-citation strength of those pages. Another way is to define the measure as a function of the confidence of the underlying association rules. For size of two, both support and confidence provide similar information. In fact, if two items A and B are presented in equal number of all transactions then there is a direct correspondence between the support and the confidence of the rules between those two items. However, support carries much less information for the frequent itemsets whose size are greater than two, as in general, the support of a large frequent itemset will be much smaller than the support of a smaller one. Furthermore, for these larger frequent itemsets, confidence of the underlying association rules can capture correlation among items that is not captured by support. We now use an example to illustrate the analysis above.

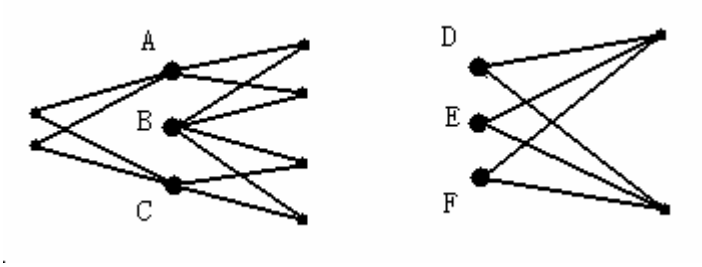


Figure 4. Two distinct page sets and their relationships

**Example 4.2** Considering page sets {A, B, C} and {D, E, F} and their relationships depicted in Figure 4, the support (co-citation strength) of each pair {A, B}, {B, C} and {C, A} is 2, and the support of {D, E}, {D, F}, {E, F} and {D, E, F} is also 2. However, the 3-way relationship among D, E, and F is much stronger than those among the pairs of {A, B}, {A, C}, and {B, C}, as *P* (D | EF), *P* (E |D F), and *P* (F | DE) are all 100%. These condition probabilities are captured in the confidence of corresponding association rules. ■

This example again illustrates that the relationships among page sets with size greater than two cannot always be captured by the pair-wise relationships. Thereby, for each frequent itemset discovered by association mining algorithms, the measure of relationships among the pages in it is defined as the average confidence of association rules, called essential rules, which have all the items of the frequent itemset and has a singleton right hand side. They are called essential rules, as they capture information unique to the given frequent itemset. This measure of relationships among pages in the frequent itemset is referred as association rule measurement.

In the mathematics formulation, given a certain frequent itemset $I$ whose confidences of essential association rules are $\{\mu_1, \mu_2 \ldots \mu_k\}$, the association rule measure of $I$ is denoted as

$$\varepsilon(I) = (\Sigma_{t=1}^{k} \mu_t) / k.$$

The extension from pair-wise pages to sets of arbitrary cardinalities means there is page-set overlap, that is, page-sets are non-disjoint. Such overlaps are not possible with pairs of pages. The measure $\varepsilon(I)$ of sets with arbitrary cardinalities is thus represented as lattices rather than $n \times n$ matrices for $n$ pages. The total number of

possible page-sets over all cardinalities is $2^n$.

In our context, the basis of representative scores computation is a weighted undirected graph whose edge represents one-order similarity of pair-wise nodes. But the measure of arbitrary cardinality sets means the similarity graph edges are generalized to hyperedges, which might involve more than two vertices. While it is a challenging work to generalize the computation of representative scores to such a similarity hypergraph, we currently implement the algorithm using a simpler hybrid way, which is between standard pair-wise similarity and the measure of sets with arbitrary cardinalities. The standard one-order similarities can be extended from the features of measure of sets with arbitrary cardinalities by summing $\varepsilon$ over all page-sets that contain the page pair involved. Formally, the new one-order similarity is defined as

$$\varsigma_I(i,j) = \Sigma_{\{I \mid i,j \in I\}} \varepsilon(I),$$

which could be easily applied to the computation of representative scores. The one-order similarity of node with itself is not redefined. The new one-order pair-wise similarity indeed eliminates many distortions of high-order similarity. The relationships of the nodes in Example 4.1 (see Figure 1) are recomputed with our new definition in Example 4.3.

**Table 2.** New one-order similarities of nodes in Figure 1

| Part | *A* | B |
|---|---|---|
| $Similarity_1^{in}(i, i)$ | 2 | 2 |
| $Similarity_1^{in}(i, k)$ | 1 | 2 |
| $Similarity_1^{in}(k, k)$ | 4 | 2 |
| $Similarity_1^{in}(k, j)$ | 1 | 2 |
| $Similarity_1^{in}(j, j)$ | 2 | 2 |

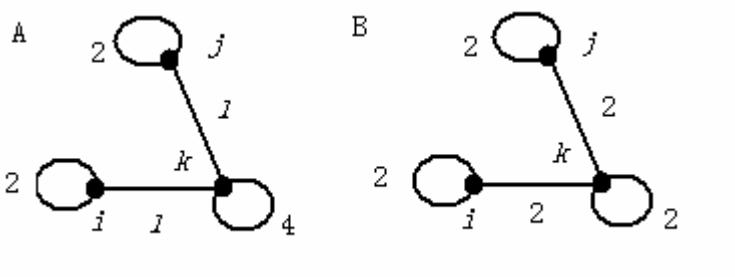


Figure 5. The path between nodes *i* and *j* in Figure 1

**Example 4.3** For the purpose of comparison, the confidences of association rules corresponding to the related sets are all set to 1. The new one-order similarities under the new definition are recomputed, and presented in the Table 2. The associated path graphs are redrawn in Figure 5, and the 5-order similarity of *i* and *j* under the new one-order similarities are recomputed as the follows,

Part A: $\varsigma_5^{in}(i, j)$

$= Similarity_1^{in}(i, i) \bullet \varsigma_1^{in}(i, k) \bullet Similarity_1^{in}(k, k) \bullet \varsigma_1^{in}(k, j) \bullet Similarity_1^{in}(j, j)$

$=2\times1\times4\times1\times2=16$,

Part B: $\varsigma_5^{in}(i,j)$

$= Similarity_1^{in}(i,i) \bullet \varsigma_1^{in}(i,k) \bullet Similarity_1^{in}(k,k) \bullet \varsigma_1^{in}(k,j) \bullet Similarity_1^{in}(j,j)$

$=2\times2\times2\times2\times2=32$.

Interestingly, the 5-order similarity of node *i* and *j* in part *B* is just twice the 5-order similarity of node *i* and *j* in part *A*. The conclusion is consistent with our initial intuitive judgment. ■

Up to now, a generalized one-order similarity is defined, under which the distortion of high-order similarity could be alleviated. In next section, the generalized similarities are applied to the proposed algorithm for topic exploration and distillation.

# 5 Similarity-based Topic Exploration and Distillation (STED) Algorithm

The proposed one-order similarities are applied to the construction of the undirected weighted graph $G_s^{in}$ and $G_s^{out}$, which are the basis of the computation of representative pages. The construction of the similarity matrix *S* of graph $G_s$ follows two steps below.

1. For the entry $S(i,j)$ of matrix *S* where *i* is equal to *j*, $S(i,j) = Similarity_1(i,j)$.
2. For the entry $S(i,j)$ of matrix *S* where *i* is not equal to *j*, $S(i,j) = \varsigma_1(i,j)$.

The iterative operation is employed in the STED algorithm. Namely, we maintain and update numerical weights for each page. Thus, with each page *p*, we associate a non-negative authority weight $x^{<p>}$ and a non-negative hub weight $y^{<p>}$. The invariant is maintained that $\Sigma(x^{<p>})^2 = 1$ and $\Sigma(y^{<p>})^2 = 1$ after each iteration, and the pages with larger *x*-values and *y*-values are viewed as being "better" representatives. [Klei98] has proved the convergence of the iteration. The iterative operation can be described as the procedure shown in Figure 6.

---

*Iteration (S, k)*

***begin procedure*** *Iteration*

*S: the matrix of the undirected weighted graph $G_s$*

*k: a natural number*

*Let z denote the n dimension vector (1, 1, 1 ... 1)*

*Set $x_0 := z$*

***for*** *i=0 to k-1*

$x_{i+1} = S \bullet x_i$

*Normalize $x_{i+1}$*

***endfor***

***return*** *($x_k$)*

***end procedure*** *Iteration*

---

Figure 6. The procedure of iterative operation

Typically, we set $k = 200$. Using this procedure, we could filter out the top *c* representative pages by ranking

the components of $x_k$ and $y_k$ respectively. Such an iterative operation based on the generalized similarity definition is effective in alleviating the problem of *topic drift* in many cases. However, the results of it are not all on topics for some queries with strong tendency to drift. Example 5.1 illustrates a weighted graph which causes a great risk on drifting for iterative procedure.

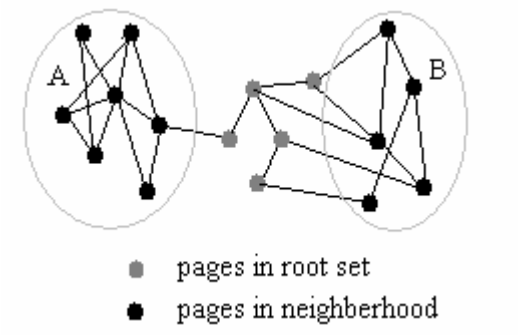


Figure 7. Weighted undirected graph $G_s$

**Example 5.1** Considering the weighted graph $G_s$ as shown in Figure 7, the top representatives returned by the iterative operation should be the pages in part A (we assign the weight of edges in Figure 7 identically to simplify the analysis), but such a result will take a great risk on drifting from initial query topic. Such a structure of weighted graph is very likely to be generated in many of the queries with strong drift tendency. Thereby, one might tend to accept the pages in part B as the result of the query since these pages seem to have more relationships with the root set. ■

Inspired from the approaches combining content analysis and connectivity analysis, the proposed definition of one-order similarity can be further extended for more effectively controlling the problem of *topic drift*. Most of the hybrid approaches using both content and topology information are based on the assumption that the pages in the root set are much less possible to drift from the query topic, which can be used as a baseline to control *topic drift* problem. Unlike those approaches using content analysis and pruning nodes from the neighborhood graph according to the baseline of root set, The STED algorithm attempts to make the definition of one-order similarity contain the baseline information. Example 4.1 illustrates the intuitive motivation and the extended definition will be presented next.

Based on the consideration above, the definition of one-order similarity is extended as follows. Given the all frequent itemsets discovered by association mining algorithms, we denote them as $I_{root}$ and $I_{neighber}$ where $I_{root}$ is the frequent itemset containing the pages in root set and $I_{neighber}$ is the frequent itemset containing no page in the root set. The one-order similarity can be defined more generally as

$$\varsigma_1(i,j) = \sum_{\{Iroot \mid i,j \in Iroot\}} \varepsilon\,(I_{root}) + \delta \cdot \sum_{\{Ineighber \mid i,j \in Ineighber\}} \varepsilon\,(I_{neighber}),$$

where $i \neq j$ and $0<\delta<1$. We call the coefficient $\delta$ is drift control factor, which reflects the discrimination extent of the relationships involving in no pages in root set. If $\delta=0$, the relationships containing no pages in root set are not token into account in the STED algorithm; if $\delta=1$, the relationships associated with any pages are considered identically. Theoretically, the smaller value of $\delta$ is assigned, the less extent of the relationships containing no pages in root set is considered. In the experiments, we found that $\delta=0$ is sufficient to return representative pages

for most of the broad topic queries. If the given query topic is rather narrow, one might consider assigning a larger value to $\delta$ in order to include more relationships into account. The STED algorithm was proved to be effective in eliminating the *topic drift* problem in most of the cases which we had encountered.

As to the problem of finding frequent itemsets $I$ and their measure $\varepsilon$ ($I$), a number of algorithms have been developed [AIS93, AS94, HS95]. *Apriori* algorithm presented in [AS94] is one of the most popular algorithms available, which can also be implemented on parallel computers [HKK97a] to use their large memory and processing power effectively. So, *Aprior* is employed in the implementation of the STED algorithm.

There are a number of settings, in which one may be interested in finding several topics relevant to the given query. As argued in section 3.2, the HITS algorithm tends to miss some less popular topics (*topic missing*) when several topics emerge in the query-specific graph. An approach as the extension of HITS was presented in [Klei98] that attempted to discover those different topics using several leading non-principal eigenvectors. Unfortunately, the experiments revealed that non-principal eigenvectors sometimes partially include the topic information, and sometimes not. And extracting the useful information from a number of non-principal eigenvectors is a hard work since each eigenvector represents a rank of all the nodes in the graph. Furthermore, unlike the principal eigenvector, the non-principal eigenvectors have both positive and negative entries, which are more difficult to make sense.

With our visualized tools, we observe that the undirected weighted graph $G_s$ tends to contain several principal connected components when the query associates with more than one topic. The assumption may be harnessed that each principal connected component in the graph $G_s$ corresponds to a potential topic associated with the given query. Thereby, a method to discover the principal topic communities appearing in $G_s$ is needed firstly instead of applying iterative operation to $G_s$ directly. Any connected components search algorithm can be applied to the STED algorithm framework for the discovery of principal topic communities. Given a *CCS* (Connected Component Search) algorithm whose input is the matrix $S$ of graph $G_s$ and output are sets of connected components $\{CC_i\}$. The topic exploration procedure is described as Figure 8.

---

*Exploration (S, τ)*
***begin procedure*** *Exploration*
*S: the matrix of the undirected weighted graph* $G_s$
*τ: a natural number*
$\{CC_i\}$ = *CCS (S)*
***for*** *i = 1 to* $|\{CC_i\}|$
  ***if*** $|CC_i| > \tau$  // $\tau$ must be set no larger than *max* ( $|CC_i|$ )
    *Construct matrix* $T_j$ *with the nodes in set* $CC_i$ *following their relationships in matrix S*
  ***endif***
***endfor***
***return*** $\{T_j\}$
***end procedure*** *Exploration*

---

Figure 8. The topic exploration procedure

The parameter $\tau$ in this procedure is the threshold which is used to drop the connected components containing nodes less than $\tau$. Such an operation is based on the assumption that the connected components containing fewer nodes in graph $G_s$ carry minorer information about query topic. One can set the threshold $\tau$ depending on the query requirement of his own. If intensive search is required, the threshold $\tau$ should be somewhat lower. However, if one intends to learn the popular information about the query, the threshold could be set higher.

The procedure *Exploration* as well as the procedure *Iteration* can be applied to topic distillation in the framework of the STED algorithm which is described in Figure 9.

In this section, we proposed a new algorithm framework of topic distillation. The proposed method retains the simple way of iterative operation in the HITS algorithm, but is better able to describe the relationships among pages thereby leading to forming a more reasonable basis to which the iterative operation applies. Furthermore, we integrate a topic exploration step in our framework before the iterative operation, and therefore enable our framework to adapt to different query requirements. We have conducted some empirical study with various queries using the STED algorithm. It is shown that the problem raised in section 3 could be almost addressed. Some of the experimental results will be discussed in the following section.

---

*Distillation ($S_{in}$, $S_{out}$, $k$, $\tau$, $c$)*

***begin procedure*** *Distillation*

*$S_{in}$: the one-order similarity matrix associated with the undirected weighted graph $G_s^{in}$*

*$S_{out}$: the one-order similarity matrix associated with the undirected weighted graph $G_s^{out}$*

*$k$, $\tau$, $c$: natural numbers*

*$\{T_j^h\}$ = Exploration ($S_{out}$, $\tau$)*

*$\{T_j^a\}$ = Exploration ($S_{in}$, $\tau$)*

***for*** *$j$ = 1 to $|\{T_j^a\}|$*

    *$x_k^j$ = Iteration ($T_j^a$, $k$)*

    *Report the pages with the $c$ largest coordinates in $x_k^j$.*

***endfor***

***for*** *$j$ = 1 to $|\{T_j^h\}|$*

    *$y_k^j$ = Iteration ($T_j^h$, $k$)*

    *Report the pages with the $c$ largest coordinates in $y_k^j$.*

***endfor***

***end procedure*** *Distillation*

---

Figure 9. The algorithm framework of STED

# 6 Experimental Results and Discussion

The STED algorithm as well as HITS was implemented in our experiments with various queries. Because of space limitations, only a representative subset of results is reported in this paper. The experimental results not included in the content of this paper can be referred in the Appendix A and B.

For the generation of the base set of pages, we follow the specification of [Klei98] described earlier. For each

of the queries, we begin by generating a root set that consists of the first 200 pages returned by HotBot [HB] on the same query. The root set is then expended to the base set by including nodes that point to, or are pointed to, by the pages in the root set. In order to keep the size of the base set manageable, for every page in the root set, we only include a fixed number of pages returned from HotBot that point to this page. The graph induced by nodes in the base set is then constructed by discovering all links among the pages in the base set, eliminating those that are between pages of the same domain.

For each query, the STED algorithm and HITS are tested on the same base set. The top eight (arbitrarily) authorities returned by both algorithms are presented when single topic appears in the query, and four authorities (arbitrarily) are returned when more than one topic emerged. For evaluation purpose, we also include a list of the URL and title (possibly abbreviated) of each page which appears in the results of each algorithm.

The performance of the different algorithms will be discussed with some examples. The first example quotes one of the experimental results appeared in [Klei98], which will be used to highlight the embarrassment of topic missing suffered by HITS using non-principal eigenvectors.

| | | |
|---|---|---|
| **(*Jaguar**)** | Authorities: principal eigenvector | **Jaguar Products** |
| 0.370 http://www2.ecst.csuchico.edu/~jschlich/Jagaur/jaguar.html | | |
| 0.347 http://www-und.ida.liu.se/~t49patsa/jserver.html | | |
| 0.292 http://tangram.informatik.uni-kl.de:8001/~rgehm/jaguar.html | | |
| 0.287 http://mcc.ac.uk/dlms/Consoles/jaguar.html | | *Jaguar products page* |
| **(*Jaguar Jaguars*)** | Authorities: $2^{nd}$ non-principal vector, positive end | **Jaguar Football** |
| 0.255 http://www.jaguarsnfl.com/ | | *Official Jacksonville Jaguar NFL Website* |
| 0.137 http://naodo.net/SportServer/football/nfl/jax.html | | *Jacksonville Jaguars Home Page* |
| 0.133 http://ao.net/~brett/jaguar/index.html | | *Brett's Jaguar Page* |
| 0.110 http://www.usatoday.com/sports/football/sfn/sfn30.htm | | *Jacksonville Jaguars* |
| **(*Jaguar Jaguars*)** | Authorities: $3^{rd}$ non-principal vector, positive end | **Jaguar Auto** |
| 0.227 http://www.jaguarvehicles.com | | *Jaguar Cars Global Home Page* |
| 0.227 http://www.collection.co.uk/ | | *The Jaguar Collection – Official Web site* |
| 0.211 http://www.moran.com/sterling/sterling.html | | |
| 0.211 http://www.coys.co.uk/ | | |

Figure 10. Example6.1A: The results of queries "*Jaguar**" and "*Jaguar Jaguars*" from [Klei98]

**Example 6.1A** The query topic *Jaguar* is a useful example which has been used by many paper including [Klei98]. The topic-representative pages were picked up by hand from several leading eigenvectors (see Figure 10). For the query topics emerging in Figure 10, the strongest collections of authoritative sources concerned the *Atari Jaguar Product*, *NFL Football Team from Jacksonville*, and the *Jaguar Automobile*. The non-principal vector selecting for multi-topics is a manual task, which depends on the given graph topology associated with the query. In Figure 10, the eigenvectors selecting of query topic *Jaguar* is based on two distinct queried topologic graphs associated with the different queries "*Jaguar**" and "*Jaguar Jaguars*" respectively: namely, *Jaguar Product* topic is derived from the query *Jaguar** while *Jaguar Mobile* and *Jaguar Football* topics are found in the query *Jaguar Jaguars*. ■

In fact, our experiments reveal that the query specific graphs of "*Jaguar**" and "*Jaguar Jaguars*" both contain the information of the three topics discovered by Example 6.1A. However, HITS failed to discover the three topics in the same query specific graph. For the purpose of comparison with our algorithm, we implement the query "*Jaguar Jaguars*" with the HITS algorithm in our experiments. The description of the results is presented in Example 6.1B.

**Example 6.1B** In our experiment, the root set of the query "*Jaguar Jaguars*" consists of 200 pages, and 1690 pages are included in the expended base set. A list of several page collections with the most positive entries of the three leading eigenvectors are shown in Figure 11. In this test case, the second and the third eigenvectors of graph also contain no negative entry.

A structural view of the weighted graph $G_s^{in}$ is shown in Figure 12, where the three main connected components are marked as part A, B and C, and part D of the graph are the rest connected components containing fewer nodes and the insular nodes in the graph $G_s^{in}$. Part A, B and C contain 290 nodes, 209 nodes and 104 nodes respectively. Part D consists of 1767 insular nodes and 21 smaller connected components whose nodes amount is in the range from 2 to 38. ■

| | |
|---|---|
| (***Jaguar Jaguars***) Authorities: principal eigenvector | |
| 0.2704 http://jacksonville.zip2.com/ | *JACKSONVILLE.COM Yellow Pages* |
| 0.2695 http://jaguars.jacksonville.com/ | *Florida Times-Union: Jaguars* |
| 0.2669 http://swell.jacksonville.com/ | *Swell Entertainment: Florida Times-Union* |
| 0.2625 http://cafe.jacksonville.com/ | *Welcome to the Florida Times-Union Online News Service* |
| (***Jaguar Jaguars***) Authorities: 2nd eigenvector | |
| 0.3062 http://www.macjag.com/ | *Fan's Site for the Jacksonville Jaguars* |
| 0.3008 http://www.paddlethetimucuan.com/getting_started.htm | *Paddling and Poking in the Timucuan* |
| 0.3008 http://www.paddlethetimucuan.com/paddle_post_subscribe.htm | |
| 0.3008 http://www.paddlethetimucuan.com/paddle01.htm | |
| (***Jaguar Jaguars***) Authorities: 3rd eigenvector | |
| 0.3016 http://www.jagnet.com/ | *JAGUAR OWNERS CLUB* |
| 0.2779 http://www.cableone.net/jcca/ | *Jaguar Club of Central Arizona* |
| 0.2244 http://www.southfloridajaguarclub.org/ | *South Florida Jaguar Club* |
| 0.2159 http://www.jagweb.com/ | *JagWeb-Jaguar restoration, trimming, bodywork, performance, etc.* |

Figure 11. Example6.1B: Our results of query "*Jaguar Jaguars*" using the HITS algorithm

It can be noticed that the three topics have emerged as part A, B, and C in Figure 12. However, the phenomenon in Example 6.1A recurred. Only two of the three topics (*NFL Football Team from Jacksonville,* and *the Jaguar Automobile*) can be picked out from Figure 11. More eigenvectors may contain further information about all topics. But extracting the valuable information from a number of non-principal eigenvectors manually is a trivial task. Furthermore, more non-principal eigenvectors computation will become less feasible when the dimension of matrix is high.

Apart from the problem of *topic missing*, one may notice that many of the returned pages of HITS in Example 6.1B are less representative or even irrelevant to the associated topics. Dot lines are used to highlight such results in Figure 11. Thereby example 6.1B also reveals the *topic drift* problem that the HITS algorithm suffers. With the

same base set of the query of "*Jaguar Jaguars*", we implement the STED algorithm and report the experimental results in the Example 6.1C.

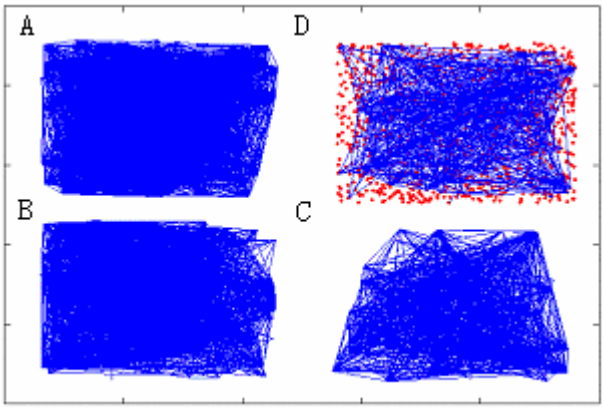


Figure 12. The topologic graph of $G_s^{in}$ of query "*Jaguar Jaguars*"

**Example 6.1C** The topologic graph of $G_s^{in}$ constructed by the STED algorithm is presented in Figure 13. Surprisedly, it contains 4 principal connected components, which present the 4 different topics including jaguar as mammal additionally. On the basis of the constructed graph $G^{in}$, the *distillation* procedure of the STED algorithm outputs the experimental results of query "*Jaguar Jaguars*" in Figure 14. ■

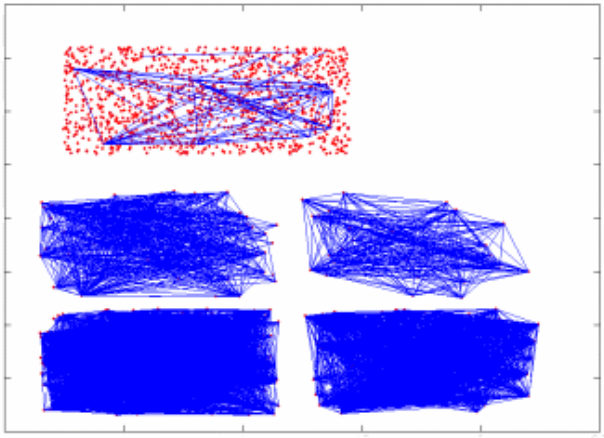

Figure 13. Topology of graph $G^{in}$ constructed by the STED algorithm

Observing the experimental results in Example 6.1C, four potential topics associated with the query are presented and each of them has its representative pages. No irrelevant pages are emerged in the results of Example6.1C. Compared with Example 6.1B, Example 6.1C illustrates the features of the STED algorithm, which is to argue the HITS algorithm with both topic drift immunity and topic exploration ability. More examples involving more than one topic can be referred in Appendix A such as the query "*Latex*". The topologic view of graph $G_s^{in}$ associated with "*Latex*" is presented in Appendix B.

In fact, the experiments of the HITS algorithm indicate that the less representative or even irrelevant results are often got not only in the situation of multi-topics involved, but many single-topic queries also tend to get irrelevant returned pages. Example 6.2A shows the case that HITS encountered in our experiments.

(***Jaguar Jaguars***) Authorities: **Jaguar Auto**

0.5399 http://www.xks.com/ *Jaguar parts and restoration, Land Rover parts*
0.3465 http://www.jag-lovers.org/ *Jag-lovers-the Jaguar Enthusiasts' premier resource on the Internet*
0.1822 http://www.bitcon.no/~gunnar/sovereign.html *Jaguar Sovereign*
0.1618 http://dmoz.org/Recreation/Autos/Makes_and_Models/Jaguar/ *Open Directory Autos: Jaguar*

(***Jaguar Jaguars***) Authorities: **NFL football team from Jacksonville**

0.4370 http://www.jax-inter.net/users/kerbob/jag.htm *The Jacksonville Jaguars NFL football team*
0.3970 http://www.macjag.com/ *Jaguars Jacksonville NFL football fan site*
0.3311 http://jaguars.jacksonville.com/ *Florida-Times Union: the ultimate Jacksonville Jaguars web site*
0.1227 http://www.jacksonville.com/tu-online/stories/072901/jag_6809002.html
*Jaguars: Brunell welcomes season of change for better 07/29/01*

(***Jaguar Jaguars***) Authorities: **Jaguar Products**

0.4771 http://www.atarihq.com/interactive/ *Atari Jaguar Discussion Board*
0.4278 http://angelfire.lycos.com/nv/jaguartop50/
0.2296 http://www.bowes.co.uk/justclaws/devcats/playtest/ *Atari Jaguar Play-Testers Register*
0.1924 http://jaguar.emugaming.com/index2.html *Atari Jaguar Front Page News*

(***Jaguar Jaguars***) Authorities: **Touring: Jaguar Reef Lodge**

0.4685 http://www.divejaguarreef.com/ *Dive Jaguar Reef Lodge, Hopkins, Belize - SCUBA diving Belize*
0.4025 http://www.jaguarreef.com/ *Jaguar Reef Lodge, Hopkins, Stann Creek, Belize, Central America*
0.3310 http://www.belizezoo.org/zoo/zoo/mammals/jag/jag1.html *The Belize Zoo - Jaguar*
0.1789 http://www.jaguarreef.com/jagreef/rates02.html *Jaguar Reef Lodge-Facilities & Rates, Belize*

Figure 14. Results of query "*Jaguar Jaguars*" using the STED algorithm

**Example 6.2A** On the query of "*Movie Awards*" using HITS, the base set of the query consists of 3026 pages. A structural view of the weighted graph $G_s^{in}$ is shown in Appendix B that contains only one principal connected component. The experimental results (see Figure 15) drift into a nepotistic clique from *searchbeat.com* community, which are irrelevant to the original query topic *Movie Awards*. Intensive study indicates that a densely connected conglomerate irrelevant to the query topic forms distorted similarities, which mislead the HITS algorithm to return the results in *seachbeat.com* community. ■

(***Movie Awards***) Authorities: principal eigenvector

0.1485 http://www.searchbeat.com/sites.htm *Site Map & Directory - Search Beat*
0.1485 http://www.searchbeat.com/featured-sites.htm *Search Beat - The One-Stop Web Guide & Directory*
0.1485 http://news.searchbeat.com/weather.htm *Weather Internet Web Links ... The Weather Beat*
0.1485 http://recreation.searchbeat.com/*traffic.htm* *Traffic and Weather Web Links ... The Traffic Beat*
0.1485 http://home.searchbeat.com/fix-it.htm *Fix It, Repair and Maintenance - The Home and Garden Beat*
0.1485 http://kids.searchbeat.com/ *The Kids & Teen Beat*
0.1485 http://history.searchbeat.com/ *History Timelines on the Web ... The History Beat*
0.1485 http://searchbeat.com/Society/Genealogy/ *Genealogy - Search Beat*

Figure 15. Result of query "*Movie Awards*" using the HITS algorithm

**Example 6.2B** Employing the STED algorithm with the same base set used by Example 6.2A, the query gets

the returned results of query "*Movie Awards*" that are listed in Figure 16, Those pages in the results could be qualified as the representative pages as for the query topic. The topologic view of the query can be referred in Appendix B. ■

(***Movie Awards***) Authorities:

| | |
|---|---|
| 0.2591 http://uk.imdb.com/Sections/Awards | *Awards & Festivals Browser* |
| 0.2588 http://reviews.imdb.com/Oscars/Awards | *Awards Season 2001 Coverage: Nominees & Winners* |
| 0.2588 http://italian.imdb.com/Oscars/Photos/Memories | *Awards Season 2001 Coverage: Oscar Memories* |
| 0.2588 http://boards.imdb.com/Guides/awards | *Movie Related Award Winners and Nominees* |
| 0.2588 http://university.imdb.com/Sections/Awards | *Awards & Festivals Browser* |
| 0.2311 http://us.imdb.com/Sections/Awards/MTV_Movie_Awards | *MTV Movie Awards* |
| 0.2298 http://www.oscars.org/ | *the Academy of Motion Picture Arts and Sciences* |
| 0.1209 http://www-scf.usc.edu/~derekj/jonathan/imdb/Pawards.html | *Awards for Jonathan Duval* |

Figure 16. Result of query "*Movie Awards*" using the STED algorithm

Example 6.2A reveals that the drifting embarrassment is also suffered by HITS on the query of single topic. The comparison between the experimental results of HITS and STED again demonstrates the topic drift immunity of the STED algorithm in the queries of single topic. Other striking examples also include the queries of "*Field Hockey*", "*Java*", "*Abduction*" etc., where the STED algorithm avoids the drift problem. But the results of HITS on those queries are all dominated by the irrelevant pages from densely connected communities. (In our ongoing work, we are getting the improvement evaluated by volunteers for comparison with similar resources in well known web directories such as Yahoo! and Infoseek, etc.)

From the overall evaluation of STED and HITS, the STED algorithm with $\delta = 1$ dramatically alleviates the problem of topic drift in HITS. However, it does not work very well for some queries with a strong drifting tendency such as the queries of *Movie Awards*, *Cancer*, etc. To control the drift more effectively, we perform the STED algorithm with different value of parameter $\delta$. The results of evaluation are shown in Figure 11. Tuning the parameter $\delta$ of drift control factor in STED, we successfully eliminate the topic drift problem in most of the cases.

In the implementation of STED, support is required as a parameter of the algorithm. Unlike the other applications of *Aprior*, the support in the STED algorithm is set very low so as to discover the frequent itemsets whose co-citation strength is more than or equal to 1. One may argue that the computation effort increases dramatically if the minimum support decreases. It is indeed the case. However, our experiments reveal that the queries of broad topics are effectively kept away from drift when $\delta = 0$. Thereby, only the frequent itemsets containing the pages in the base set are needed to discover. Such a constraint dramatically alleviates the computation complexity of computing frequent itemsets caused by lower minimum support. However, dropping relationships containing no page in root set might deteriorate the quality of the returned pages when the query topic is rather narrow. Fortunately, when the query topic is narrow, the computation of finding frequent itemsets in fact is much less complex even if the relationships among all the pages are to be considered. The reason is obvious, that is, there are much fewer relationships among the all pages when the query topic is narrow.

The experiments indicate the difference between the behavior of the HITS algorithm and STED. In particular, when computing the top authorities, the HITS algorithm tends to concentrate on the most popular topic (densely connected community) in the query specific graph, whereas our algorithm will take also less popular topics into

account and tends to return pages from different topics with classified results.

The effect of "densely connected community" in HITS may become poignant in many of the cases where the densely connected community has little or nothing to do with the proposed query topic. A more elaborate algorithm for weighting links such as [BH98] could help alleviate this problem. However, our experiments seem indicative of the topic drift potential of the HITS algorithm, which has been studied theoretically in section 3.

## 7 Conclusion and Future Work

In this paper, we showed the intrinsic reason of topic drift potential in HITS-like algorithms by a similarity-based analysis model. To address the drift problem, we presented a methodology for topic distillation that integrates association mining techniques into regular hyperlinks analysis. The proposed method achieves a considerable improvement in precision without employing any textual information. Furthermore, we integrate a topic exploration component into our algorithm framework, which enables the end-users to search less popular topic on the Web. To our knowledge, the follows are the first time to be proposed:

1. A similarity-base analysis model applied to topic distillation.
2. The application of association mining integrated into the hyperlink analysis for the purpose of topic distillation.
3. A topic exploration function in distillation algorithm.

In our ongoing work, apart from completing a detailed user study, we are exploring two more ideas. First, deriving the representatives from the hypergraph of arbitrary cardinality similarity is an interesting work. Second, a more refined similarity definition can be expected to obtain representatives in one step instead of a number of iterations.

## References


[AIS93] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. *Proc. of ACM-SIGMOD'93 Int. Conf. on Management of Data*, 1993.

[AS94] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. *Proc. of the 20th VLDB Conference*, pages 487-499, 1994.

[BH98] K. Bharat and M. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. *Proc. of 21st ACM-SIGIR Int. Conference on Research and Development in Information Retrieval*, pages 104-111, 1998.

[Bot93] R. A. Botafago. Cluster analysis for hypertext systems. *Proc. Of 16th ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 116-125, 1993.

[BP98] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Proc. of 7thACM-WWW Int. Conference*, 1998.

[BRT01] A. Borodin, G. Roberts, J. Rosenthal and P. Tsaparas. Finding authorities and hubs from link structures on the World Wide Web. *Proc. of 10thACM-WWW Int. Conference*, 2001.

[CC98] C. Chen and M. Czerwinski. From latent semantics to spatial hypertext – An integrated approach. *Proc. of the 9th ACM Conference on Hypertext and Hypermedia*, pages 77--86, 1998.

[CDG98] S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, P. Raghavan and S. Rajagopalan. Automatic resource compilation by analyzing hyperlink structure and associated text. *Proc. of 7th World Wide Web Conference*, pages 65-74, 1998.

[Cha01] S. Chakrabarti. Integrating the document object model with hyperlinks for Enhanced topic distillation and information extraction. *Proc. of 10thACM-WWW Int. Conference*, 2001.

[Che97] C. Chen. Structuring and visualizing the WWW by generalized similarity analysis. *Proc. of ACM-Hypertext '97 Conference*, 1997.

[DGK00] B. Davison, A. Gerasoulis, K. Kleisouris, Y. Lu, H. Seo, W. Wang, and B. Wu. DiscoWeb: Applying link analysis to web search (extended abstract). *Poster proceedings of the Eighth International World Wide Web Conference***,** pages 148-149, 1999.

[GL89] G. H. Golub and C. F. van Loan. Matrix Computations. *Johns Hopkins University Press*, London, 1989.

[Ha75] J. Hartigan. Clustering algorithms. *John Wiley & Sons*, New York, 1975.

[HB] http://www.hotbot.com

[Hea97] M. Hearst. Distinguishing between web data mining and information access: Position statement. *KDD'97 Panel on Web Data Mining*, 1997.

[HKK97a] E. H. Han, G. Karypis, V. Kumar. Scalable parallel data mining for association rules. *Proc. of ACM-SIGMOD Int. Conf. on Management of Data*, 1997.

[HKK97b] E. H. Han, G. Karypis, V. Kumar, and B. Mobasher. Clustering based on association rule hypergraph. *Proc. of the Workshop on Research Issues on Data Mining and Knowledge Discovery*, pages 9-13, 1997.

[HS95] M. A. W. Houtsma and A. N. Swami. Set-oriented mining for association rules in relational databases. *Proc. of the 11th Int'l Conf. on Data Eng.*, pages 25-33, 1995.

[Klei98] J. Kleinberg. Authoritative sources in a hyperlinked environment. *Proc. of 9th ACM-SIAM Symposium on Discrete Algorithms*. Also appeared as IBM Research Report RJ 10076, May 1997.

[Lar96] R. Larson. Bibliometrics of the World Wide Web: an exploratory analysis of the intellectual structure of cyberspace. *Ann. Meeting of the American Soc. Info. Science*, 1996.

[PPR96] P. Pirolli, J. Pitkow and R. Rao. Silk from sow's ear: Extracting useable structures from the web. *Proc. of CHI'96 Conference*, pages 118-125, 1996.

[SG74] H. Small and B. Griffith. The structure of scientific literatures, I: Identifying and graphing specialties. *Science Studies*, 4(17), pages 17-40, 1974.

[Sm73] Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of American Social Information Science*, pages 265-269, 1973.

[WM89] H. White and K. McCain. Bibliometrics. *Annual Review of information Science and Technology 24*, pages 119-186, 1989.

[WM98] H. White and K. McCain. Visualizing a discipline: An author co-citation analysis of information science, 1972-1995. *Journal of the American Society for information Science, 49*, pages 327-356,1998.

# Appendix A

**Query:** *Abduction* (Base Set size = 2332)

| HITS Algorithm | STED Algorithm |
|---|---|
| http://about.com/tvradio/ | http://users.olynet.com/mkathj/MissingChildren.html<br>*Missing Children / Parental Abduction* |
| http://about.com/travel/ | http://personalwebs.myriad.net/steveb/abduct.html<br>*Child abduction, protecting your child, child abuse* |
| http://about.com/teens/ | http://pages.tca.net/steveb/abduct.html<br>*Child abduction, protecting your child, child abuse* |
| http://about.com/style/ | http://www.maplesearch.com/…Issues/Children/Child_Abduction<br>*Society/Issues/Children/Child_Abduction* |
| http://about.com/sports/ | http://www.directory.bluewindow.ch/…/Child_Abduction<br>*Child Abduction* |
| http://about.com/smallbusiness/ | http://www.webwombat.com.au/wwdir/ww29158.htm<br>*Society: Issues: Children: Child_Abduction* |
| http://about.com/shopping/ | http://www.ability.org.uk/child_abduction.html<br>*Child Abduction* |
| http://about.com/science/ | http://www.fullwebinfo.com//Top/Society/…/Child_Abduction<br>*Issues: Children, Youth and Family: Child Abduction* |
| Annotation: about.com--*network of sites by subject specialists writing articles, publishing free email newsletters and providing personally-reviewed links for each topic*. | |

**Query:** *Computation Complex* (Base Set size = 2122)

| HITS Algorithm | STED Algorithm |
|---|---|
| http://www.briansbooks.com/ | http://compgeom.cs.uiuc.edu/…/journals.html<br>*Computational Geometry Journals* |
| http://www.briansbooks.com/…/basket | http://math-www.uni-paderborn.de/~aggathen/cc<br>*Homepage of computational complexity* |
| http://www.briansbooks.com/…/privacy | http://www.informatik.uni-trier.de/~ley/db/journals/cc<br>*DBLP computational complexity* |
| http://www.briansbooks.com/…/news | http://sunsite.informatik.rwth-aachen.de/dblp/db/journals/cc<br>*DBLP computational complexity* |
| http://www.briansbooks.com/…/about | http://www.ams.org/mathweb/mi-journals5.html<br>*Math on the Web: Journals* |
| http://www.briansbooks.com/…/contact | http://math.uni-heidelberg.de/logic/bb/bblinks.html |

| | |
|---|---|
| | *Internet Links in Theoretical Computer Science* |
| http://www.briansbooks.com/…/FAQ | http://theory.lcs.mit.edu/~prahladh/bookmarks.html<br>*Bookmarks for Prahladh Harsha* |
| http://www.briansbooks.com/…/policies | http://bweb.nus.edu.sg/libweb/asp/ejsubcsm.asp<br>*Computer Science & Mathematics* |
| Annotation: http://www.briansbooks.com/ -- *Offers a great selection of internet and computer books at discounted prices every day* | |

**Query:** *Field Hockey* (Base Set size = 2705)

| HITS. Algorithm | STED Algorithm |
|---|---|
| http://www.studentadvantage.com<br>*Welcome to Student Advantage* | http://umterps.fansonly.com/sports/w-fieldh/md-w-fieldh-body.html<br>*Official Athletic Site of the University of Maryland- Field Hockey* |
| http://www.fansonly.com/channels/site/privacy.html<br>*FANSonly - Your Ticket to College Sports* | http://wakeforestsports.fansonly.com/sports/w-fieldh/wake-w-fieldh-body.html<br>*Official Athletic Site, Wake Forest Demon Deacons - Field Hockey* |
| http://www.fansonly.com/oas/<br>*FANSonly - Your Ticket to College Sports* | http://www.msuspartans.com/sports/w-fieldh/msu-w-fieldh-body.html<br>*Official Athletic Site, Michigan State Field Hockey* |
| http://www.fansonly.com/channels/news/sports/m-footbl/2001footballpreview.html<br>*FANSonly - Your Ticket to College Sports* | http://www.goblackbears.com/fieldhoc<br>*Uoflsports.com: The Official Web Site of Louisville Field Hockey* |
| http://bbs.fansonly.com/cgi-bin/wwwthreads/index.cgi<br>*FANSonly - Your Ticket to College Sports* | http://www.ohiostatebuckeyes.com/sports/w-fieldh/osu-w-fieldh-frame.html<br>*The Official Web Site of Ohio State Field Hockey* |
| http://www.fansonly.com<br>*FANSonly - Your Ticket to College Sports* | http://tarheelblue.fansonly.com/sports/w-fieldh/unc-w-fieldh-body.html<br>*Official Athletics Site of the University of North Carolina Tar Heels - Field Hockey* |
| http://www.fansonly.com/channels/news<br>*FANSonly - Your Ticket to College Sports* | http://friars.fansonly.com/sports/w-fieldh/archive/prov-w-fieldh-ncaadiv1.html<br>*Providence College Friars Official Athletic Site -Field Hockey* |
| http://www.fansonly.com/channels/news/sports/w-baskbl/polls.html<br>*FANSonly - Your Ticket to College Sports* | http://nusports.fansonly.com/sports/w-fieldh/nw-w-fieldh-body.html<br>*Wildcat Field Hockey Northwestern Wildcats-Official Athletic* |

| | *Site* |
|---|---|

**Query:** *Randomized Algorithm* (Base Set size = 1144)

| HITS Algorithm | STED Algorithm |
|---|---|
| http://www.csee.umbc.edu/~germida/algorithm_links.html<br>*Algorithms Animations* | http://www.fi.muni.cz/mfcs98/<br>*MFCS'98 home page* |
| http://www.cis.temple.edu/courses-alg.html<br>*A List of Courses on Algorithms* | http://www.cis.temple.edu/courses-alg.html<br>*A List of Courses on Algorithms* |
| http://www.loria.fr/~hermann/publications.html<br>*PUBLICATIONS* | http://www.cs.cmu.edu/afs/cs.cmu.edu/user/avrim/www/Randalgs97/home.html<br>*15-852 RANDOMIZED ALGORITHMS* |
| http://www.cs.duke.edu/~reif/Bookmarks.html<br>*Bookmarks for Conferences* | http://www.ics.uci.edu/~eppstein/<br>*Fundamental Algorithms Home Page* |
| http://www.di.unipi.it/~pisanti/bookmarks.html<br>*Bookmarks for Nadia Pisanti* | http://www.eccc.uni-trier.de/eccc/<br>*The Electronic Colloquium on Computational Complexity* |
| http://www.tuwien.ac.at/theoinf/AofA/Resource<br>*ANALYSIS of ALGORITHMS* | http://www.cs.sunysb.edu/~skiena/<br>*Fundamental Algorithms Home Page* |
| http://www.informatik.fernuni-hagen.de/cca/cca98.html<br>*CCA Net - CCA'98* | http://www.mathematik.uni-osnabrueck.de/research/OR/class<br>*Complexity results for scheduling problems* |
| http://www.dbai.tuwien.ac.at/staff/gottlob/<br>*DBAI -- Prof. Dr. Georg Gottlob* | http://www.nada.kth.se/nada/theory/aalg/lectures.html<br>*Lecture notes Advanced Algorithms* |

**Query:** *Java* (Base Set size = 1725)

| HITS Algorithm | STED Algorithm |
|---|---|
| http://www.internet.com/corporate/legal.html<br>*Welcome to internet.com's International Channel* | http://www.scriptsearch.com/<br>*Ultimate Resource for scripts, source code* |
| http://www.internet.com/corporate/privacy/privacypolicy.html<br>*INT Media Group, Incorporated Privacy Policy* | http://www.webdeveloper.com/<br>*Web Developers and Designers Learn How to Build Web Sites, Program in Java and JavaScript* |
| http://www.internet.com/sections/resources.html<br>*Welcome to internet.com's Internet Resources Channel* | http://javascript.internet.com/<br>*Free JavaScripts, Tutorials, Example Code, Reference, Resources, And Help* |
| http://www.internet.com/sections/isp.html<br>*Welcome to internet.com's ISP Resources Channel* | http://www.sun.com/java/<br>*Free Java Applets, Games, Programming Tutorials, and Downloads* |

| | |
|---|---|
| http://www.internet.com/sections/downloads.html<br>*Welcome to internet.com's Download Channel* | http://redir.internet.com/<br>*Internet.com redirection server* |
| http://www.internet.com/sections/linux.html<br>*Welcome to internet.com's Linux/Open Source Channel* | http://www.htmlgoodies.com/<br>*HTML Goodies – Home Page* |
| http://www.internet.com/sections/lists.html<br>*Welcome to internet.com's Internet Lists Channel* | http://softwaredev.earthweb.com/java<br>*EarthWeb.com: The IT Industry Portal* |
| http://www.internet.com/sections/webdev.html<br>*Welcome to internet.com's Web Developer Channel* | http://www.jars.com/<br>*The #1 Java Review Service* |

**Query:** *Latex* (Base Set size = 2315)

| HITS Algorithm | STED Algorithm | |
|---|---|---|
| http://www.giss.nasa.gov/latex/<br>hypertext Help with LaTeX | Topic1 | http://www.tug.org/<br>TeX Users Group Home Page |
| http://www.giss.nasa.gov./index.html<br>NASA Goddard Institute for Space Studies | | http://www.giss.nasa.gov/latex/<br>hypertext Help with LaTeX |
| http://www.giss.nasa.gov/latex/ltx-2.html<br>Help On LaTeX Commands | | http://www.maths.tcd.ie/~dwilkins/LaTeXPrimer/<br>Getting Started with LaTeX |
| http://www.giss.nasa.gov/latex/externals.html<br>Other On-line LaTeX Information | | http://www.latex-project.org/intro.html<br>An Intrduction to LaTeX |
| http://www.giss.nasa.gov/latex/ltx-tar.html<br>Hypertext LaTeX Help files | Topic2 | http://www.familyvillage.wisc.edu/lib_latx.htm<br>The Family Village / Libraby / Latex Allergy |
| http://www.giss.nasa.gov/latex/refs.html<br>LaTeX Reference Books | | http://www.notvanilla.com/gayscape/rubber.html<br>Rubber and Latex by Gayscape |
| http://www.giss.nasa.gov/latex<br>/LaTeX-info.html<br>About Help On LaTeX | | http://dir.yahoo.com/Health/Diseases_and_Conditions<br>/Latex_Allergies/<br>Yahoo! Health > Diseases and Conditions > Latex Allergies |
| http://www.giss.nasa.gov/more/<br>NASA Goddard Institute: More Resources | | http://www.sinuses.com/allergy.htm<br>Allergy - Sinusitis - WS Tichenor MD |

**Query:** *Table Tennis* (Base Set size = 2906)

| HITS Algorithm | STED Algorithm |
|---|---|
| http://about.com/movies/ | http://tabletennis.about.com/mbody.htm |
| http://about.com/musicperform/ | http://tabletennis.about.com/bl-webring.htm |
| http://about.com/newsissues/ | http://tabletennis.about.com/ |
| http://about.com/parenting/ | http://tabletennis.miningco.com |
| http://about.com/people/ | http://tabletennis.about.com/sports/recreation/tabletennis/ |

| | |
|---|---|
| http://about.com/pets/ | http://tabletennis.about.com/sports/tabletennis/cs/coaching/index.htm |
| http://about.com/recreation/ | http://tabletennis.about.com/sports/tabletennis/cs/rules/index.htm |
| http://about.com/realestate/ | http://tabletennis.about.com/sports/tabletennis/library/weekly/topicmenu.htm |

## Appendix B

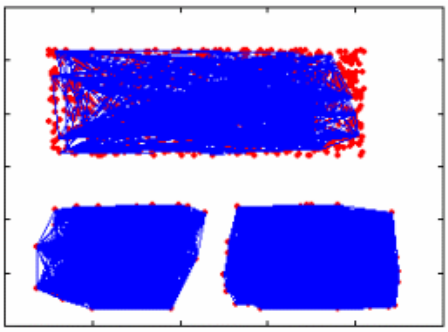

Figure $A_1$

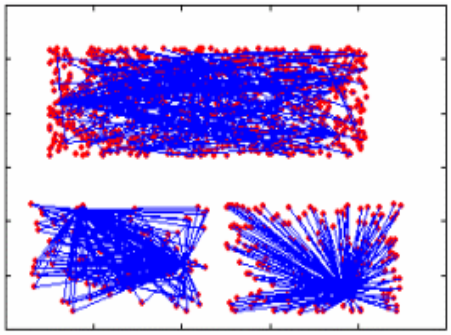

Figure $A_2$

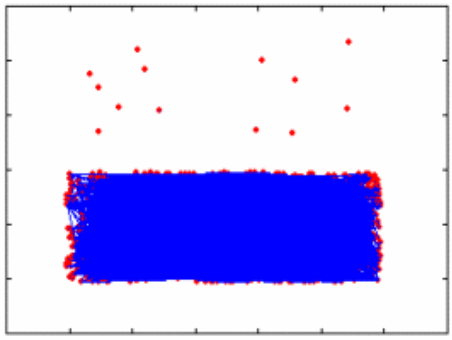

Figure $B_1$

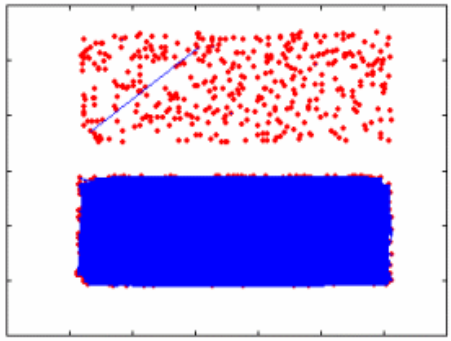

Figure $B_2$

Figure $A_1$. Topology of graph $G^{in}$ of query "*Latex*" constructed by the HITS algorithm

Figure $A_2$. Topology of graph $G^{in}$ of query "*Latex*" constructed by the STED algorithm

Figure $B_1$. Topology of graph $G^{in}$ of query "*Movie Awards*" constructed by the HITS algorithm

Figure $B_2$. Topology of graph $G^{in}$ of query "*Movie Awards*" constructed by the STED algorithm