Soft Comput (2006) 10: 2–11
DOI 10.1007/s00500-005-0457-0

FOCUS

Zhuoming Xu · Xiao Cao · Yisheng Dong
Yahong Han

s-HITSc: an improved model and algorithm for topic distillation on the Web

Published online: 15 June 2005

Abstract Topic distillation on the Web, namely, finding quality information sources related to a given query topic with hyperlink analysis, has been shown to be useful in Web IR. Based on the analysis of three deficiencies of classical topic distillation algorithm HITS, this paper presents an improved model and algorithm named s-HITSc. Given a query topic, the improved algorithm can model a neighborhood graph at site granularity, compute the relevance weights of the nodes to the topic with content analysis, and apply weighted I/O operations in its iterative hyperlink analysis. Theoretical analysis and experimental results show that s-HITSc can control topic drift and identify more reasonable and meaningful authority and hub sites on a given topic.

Keywords Topic distillation · Site granularity · Hyperlink analysis · Content analysis · Web IR

1 Introduction

For many topics, the Web contains thousands of relevant information sources of widely varying quality. The users, who make topic searches on the Web, face a daunting challenge in identifying a small subset of quality topical sources worthy of their attention, especially when using traditional text-based search tools. Since the Web is a hyperlinked environment, *hyperlink analysis* algorithms are provenly able to significantly improve the quality of search results [10].

As a hyperlink analysis algorithm, *topic distillation* is a process of finding quality information sources related to a given query topic by analyzing the structure of a query-specific link graph, called a *neighborhood graph* [13, 10]. Kleinberg's Hypertext Induced Topic Search (HITS) algorithm [13] is the first and best-known topic distillation algorithm, which has been implemented by IBM's Clever project [5]. However, many further researches (such as [3], [15] and [4]) and our investigation show that there exists several problems when HITS algorithm is applied to practice. The problems include *failing to meet users' site-granularity information needs*, *tending to produce unreasonable results*, and *topic drift* (i.e., the top-ranked information sources are not (completely) on the original query topic). Based on the analysis of these deficiencies of HITS algorithm, this paper presents an improved topic distillation model and algorithm named s-HITSc (*s*ite-granularity HITS enhanced by *c*ontent analysis). Theoretical analysis and experimental results show that our new algorithm can control topic drift and identify more reasonable and meaningful Web sites on a given query topic.

The rest of the paper is organized as follows. Section 2 recounts Kleinberg's HITS Algorithm, and enumerates the problems identified by our investigation and other related researches. Section 3 presents our site-granularity topic distillation model and algorithms that tackle the problems. In Sect. 4, we make further theoretical analysis to indicate that the iterative procedure of our improved algorithm converges and produces more reasonable and meaningful results. Section 5 gives our experimental results to show the effectiveness of the improved algorithm. Finally, we conclude our work.

2 Related work and motivation

Despite the decentralized and unorganized nature of the Web, many studies [14] have shown that the Web has

Z. Xu (✉) · Y. Dong
Department of Computer Science and Engineering,
Southeast University, Nanjing, 210096, China
E-mail: zmxu@acm.org
E-mail: ysdong@seu.edu.cn

Z. Xu · X. Cao · Y. Han
College of Computers and Information Engineering,
Hohai University, Nanjing, 210098, China
E-mail: xcao@ieee.org
E-mail: hanyahong@hotmail.com

self-organization via hyperlinks among information sources (Web pages or sites). At a global level, the information sources on a common topic tend to cluster into a natural *community*. At a local level, there exists a characteristic pattern: A community consists of a collection of *authorities* (sources with quality content on the topic) linked in a correlated fashion by a collection of *hubs* (guides and resource lists with many hyperlinks to authorities).

From the viewpoint of IR, the notion of authority adds a crucial second dimension to the concept of *relevance*: Users wish to search not only a set of relevant sources, but also those relevant sources of the highest *quality* [5]. The self-organization of the Web provides the foundation for topic distillation. As a *connectivity-based ranking*, topic distillation algorithms make the following simplifying assumptions [10]:

- *Assumption 1* (*quality presumption*). A hyperlink from page A to page B is a recommendation of page B by the author of page A.
- *Assumption 2* (*relevance presumption*). If page A and page B are connected by a hyperlink, then they might be on the same topic.

2.1 Kleinberg's HITS algorithm

Topic distillation adopts a *query-dependent* scheme. Given a user query, the basic idea of HITS algorithm is to build a query-specific link graph, called a *(page) neighborhood graph*, and perform hyperlink analysis on the graph to identify top authorities and hubs. The graph is built as follows [13]:

1. A *root set* R_Q of pages matching the query Q is collected from a text-based search engine or text index system (say, the top 200 items).
2. R_Q is augmented by its neighboring pages, which are a limited number of pages (say, 50) that hyperlink to pages in R_Q, and all pages that are hyperlinked to by pages in R_Q. The root-set pages and the neighboring pages form a *base set* B_Q.
3. A *neighborhood graph* $G[B_Q]$ is induced by B_Q, where each page in B_Q is modeled by a node, there exists a directed edge from node A to node B iff page A hyperlinks to page B.

To identify good authorities and hubs, the algorithm computes an *authority score* and a *hub score* for each node in graph $G[B_Q]$, and then ranks the nodes by those scores. Obviously, authorities and hubs exhibit *mutually reinforcing (MR)* relationship: A good authority is a page that is hyperlinked to by many good hubs; a good hub is a page that hyperlinks to many good authorities. This motivates the definition of two operations ***I*** and ***O*** on the authority and hub scores, respectively [13]. Given a neighborhood graph $G[B_Q]=(V, E)$, for each node $u \in V$, the authority score $\mathbf{a}[u]$ and hub score $\mathbf{h}[u]$ of node u are computed by:

I operation: $\mathbf{a}[u] := \sum_{\{v|(v,u)\in E\}} \mathbf{h}[v]$, and
O operation: $\mathbf{h}[u] := \sum_{\{v|(u,v)\in E\}} \mathbf{a}[v]$.

To break the circularity when computing vector $\mathbf{a} \in R^{|V|}$ and $\mathbf{h} \in R^{|V|}$, HITS uses the following iterative procedure [13]:

1. Initialize vector $\mathbf{a}$ and $\mathbf{h}$ to $(1, 1, \ldots, 1)^{\mathrm{T}} \in R^{|V|}$.
2. While vector $\mathbf{a}$ and $\mathbf{h}$ have not converged, do:

(a) Apply the ***I*** operation on $\mathbf{h}$ to compute $\mathbf{a}$.
(b) Apply the ***O*** operation on $\mathbf{a}$ to compute $\mathbf{h}$.
(c) Normalize $\mathbf{a}$ and $\mathbf{h}$.

3. Report top 5–10 authorities and hubs.

2.2 Problems and analysis

Our investigation and other related researches show that there exist several problems when HITS algorithm is applied to practice. We enumerate the problems and analyze them as follows.

Problem 1. *Failing to meet users' site-granularity information needs*. All existing topic distillation algorithms (including HITS) model the graph at *page granularity* and perform the analysis at a page level. Bharat and Henzinger [3] pointed out: hypertext encourages documents to be split up into pieces. What users are looking for on the Web are good *sites*, containing a set of connected pages on the topic, rather than individual pages. The analysis of AltaVista's transaction logs [6] showed that many users often raise topic-specific site finding queries to the search engine. To meet users' site-granularity information needs, many today's search engines (such as Google, see [17] for more) can deliver retrieved results at *site granularity*—all the pages (URLs) coming from the same site (domain or host) cluster together as a single item in the result list. Some of them have provided "Top 10 Results (pages)" services, but have not provided similar services at site granularity (at least no any published algorithm exists, to the best of our knowledge).

Problem 2. *Tending to produce unreasonable results*. As mentioned earlier, topic distillation is the process of information (re-)filtering based on a *quality* metric (authority). The quality of an information source inherently is a matter of human judgement [1]. Therefore, the rationale of topic distillation is that *we gain a common understanding of the quality of information sources by mining the collective quality-judgement made by different creators of pages and hyperlinks (i.e., the authors of Web sites)*. Since all pages and the hyperlinks they contain on a Web site are generally considered as being created by a single author (person or group) [3, 1, 6], they should own a single share of quality-judgement. However, HITS, as a *page-granularity* algorithm (even though all intrinsic links on a site are eliminated from the graph, as suggested by [13] and [5]), tends to define unjust *influence weights* for different authors of Web sites since the neighborhood graph models different numbers

of pages and hyperlinks created by different authors, which is a deviation from the rationale of topic distillation. Worse of all, if multiple pages on site A hyperlink to a single page on site B, or the reverse case, one page on site A hyperlinks to multiple pages on site B, or both the cases (these may be prevalent clique manipulations of search engine rankings—one of the causes of why HITS algorithm is not currently used in a commercial search product [10]), then the hub scores of the page(s) on site A and the authority score of the page(s) on site B are drove up unreasonably.

Problem 3. *Topic drift* [3, 15, 4]— *the top-ranked authorities/hubs are not (completely) on the original query topic*. This is another main problem of HITS algorithm which impedes its use in practice [10]. As a MR approach, the iteration of HITS algorithm is likely to converge at a *tightly knit community (TKC)* (a highly interconnected set of nodes) on the neighborhood graph [15]. If the pages in the TKC are not or less relevant to the query topic, or the TKC is small and therefore cannot stand for the dominant topic of the base-set, then the *TKC effect* [15] occurs, thus leads to topic drift. The root of the problem is that the algorithm is cocksure of the *relevance presumption* (*Assumption 2*), and, when computing the ranking scores, gives equivalent weights to all the pages (including the augmented irrelevant pages) in the base-set—this actually means that it gives undue weights to the nodes in the (irrelevant) TKC. Based on lots of experiments, Borodin et al. [4] conclude that as a "links-only" approach, topic distillation algorithms cannot completely avoid topic drift unless they combining *content analysis* with hyperlink analysis.

3 Proposed model and algorithms

3.1 Road map for improvements

To tackle the above problems and design our new algorithm s-HITSc, we try to improve original HITS algorithm along two dimensions:

- *Modeling the neighborhood-graph at site granularity*. Viewing the Web as a site graph (see Definition 3 below) and performing hyperlink analysis on it, the new algorithm not only meets users' *site-granularity* information needs, but also avoids defining unjust *influence weights* for different authors of information sources (Web sites), thus make the results more reasonable. This is actually a reasonable amendment to the *quality presumption*.
- *Combining content analysis with hyperlink analysis*. With content analysis, the new algorithm computes the *relevance weights* of the nodes to the query topic. The weights control the computation of the authority and hub scores during hyperlink analysis, thus make the top-ranked authorities and hubs more meaningful to the topic. This is actually a reasonable amendment to the *relevance presumption*.

3.2 Aggregating pages into sites

Informally, we adopt the definition of the term "site" by [1] and [6]: A *site* (multimedia document) is an organized collection of pages (URLs) on a specific topic maintained by a single person or group. ...A site is not the same thing as a Web domain or host. ...An important feature of a site is that its pages are published together, as a coherent source of information. ...A site can be identified with a selected page (URL) in its page collection.

We now define some terms formally in the context of *site-granularity topic distillation* as follows.

Definition 1 (site s_Q): A *site* $s_Q = (id, C)$ is a collection C of pages (URLs) published on a Web site[1] that "match" the user query Q. One of the pages with the shallowest URL path depth among all the URLs of the pages in C is selected as the identifying page *id* of the site (normally, *id* is the "root page" [1] or "entry page" [6] of the site). We note that each site's page collection consists of all the corresponding pages in base-set B_Q that belong to the same Web site.

Definition 2 (site base-set $\mathcal{B}_Q$): A *site base-set* $\mathcal{B}_Q = \{s_Q^1, s_Q^2, \ldots, s_Q^m\}$ is a finite set consisting of m sites $s_Q^1, s_Q^2, \ldots, s_Q^m$. We note that conceptually, a site base-set is equivalent to a page base-set in the context of page-granularity topic distillation. But, rather than by expanding from a "root-set", it can be directly formed by aggregating the pages (URLs) in base-set B_Q.

Definition 3 (site neighborhood-graph $G[\mathcal{B}_Q]$): A *site neighborhood-graph* $G[\mathcal{B}_Q] = (V, E)$ is a directed graph induced by a site base-set $\mathcal{B}_Q$, where, each site in $\mathcal{B}_Q$ is modeled by a node, and there exists an edge from node A to node B iff there exists a page in site A's page collection hyperlinks to a page in site B's page collection.

Using simple heuristics for URL pattern analysis, the following algorithm **Aggregation** can aggregate a page base-set B_Q into a site base-set $\mathcal{B}_Q$, where function **URL_host**() returns the host to which a page belongs, function **URL_depth**() calculates the URL path depth of a page: Table a

3.3 Weighting nodes with content analysis

Early studies [3, 5] on topic distillation at *page granularity* have tried to avoid unreasonable results due to the "links-only" approach in the score computation by considering the factor of the *relevance* of the pages on the query topic. In Web search, user queries are usually *broad-topic queries* with few keywords [13, 12], therefore, just as Bharat and Henzinger [3] stated, matching the query itself against the pages is usually not sufficient. Instead, [3] use the concatenation of the first 1,000 words from each page in the root-set as a *pseudo-document* to

[1]Here, we separate Web sites by *host* (not *domain*). But, in fact, our algorithms can be applicable for both *host* and *domain* cases.

Algorithm **Aggregation**(B_Q)
//Algorithm for forming a site base-set.
$B_Q = \{p_1, p_2, \cdots, p_n\}$: a page base-set containing n pages.
1. $m := 1$; $C_1 := \{p_1\}$; $id_1 := p_1$.
2. FOR $i := 2$ TO n DO
3. [GET a page $p_i \in B_Q$, and set *put* := *FALSE*.
4. FOR $j := 1$ TO m DO
5. IF **URL_host**(p_i) = **URL_host**(id_j) THEN
6. [$C_j := C_j \cup \{p_i\}$.
7. IF **URL_depth**(p_i) < **URL_depth**(id_j) THEN $id_j := p_i$.
8. Set *put* := *TRUE*, then EXIT from the j loop.].
9. IF *NOT put* THEN [$m := m+1$; $C_m := \{p_i\}$; $id_m := p_i$.].].
10. Set sites $s_Q^1 := (id_1, C_1)$, $s_Q^2 := (id_2, C_2)$, …, $s_Q^m := (id_m, C_m)$, and site base-set $\mathcal{B}_Q := \{s_Q^1, s_Q^2, ..., s_Q^m\}$.
11. RETURN $\mathcal{B}_Q$.

redefine the query topic, and then compute the relevance weights of each page in the base-set to the topic. But text extraction and concatenation from mass pages and the redoing of lexical analysis and term-weight computation mean a high computational cost. Chakrabarti, et al. [5] use the number of matches between the query keyword(s) and the anchor text (including some surrounded words) of hyperlinks to weight the hyperlinked pages. This approach also has disadvantages: (1) Match counting between the terms (especially when the query-string contains multiple keywords) means a high computational complexity and cost; (2) it cannot weight hub pages containing only out-links; (3) it cannot deal with the synonym problem.

Most text-based Web IR systems adopt the *vector space model (VSM)* or its variants [2] as their retrieval models. Therefore, our approach is: Firstly, the query topic and the content topic of each node (site) on the site neighborhood-graph are represented as vectors in VSM. Then we compute the *cosine similarity* [2] between the query-topic vector and the node-topic vector, with which weights each node on the graph (the neighborhood graph is now transformed into a *node-weighted graph*).

We use the matched documents to redefine the query topic. This approach is commonly adopted in IR, such as in *user relevance feedback* and *query expansion* in traditional IR, and *similar-page queries* in Web search engines [2].

Baeza-Yates et al. [2] give an incisive explanation to the essence of VSM: a user query defines a fuzzy set of documents relevant to the query topic, IR problem can be reduced to the problem of *document clustering* in the term vector space, i.e. to partition all documents into two *clusters*—one belongs to the fuzzy set (i.e. the query results), the other does not. Based on this understanding and the basic principle of document clustering [11], we can represent a cluster by the *centroid* of its elements. In another words, we can represent the query topic by the centroid of the VSM vectors of all pages in the root-set.

Suppose the VSM vector of page $p_j \in R_Q$ is denoted by $\mathbf{vector}(p_j) = (w_{1j}, w_{2j}, \ldots, w_{tj}) \in R^t$, where, t is the total number of index terms in the IR system, $w_{ij} = TF_{ij} \times IDF_i$ denotes the weight of p_j on index term K_i, then the topic $\mathbf{t}_Q$ of query Q can be represented as follows:

$$\mathbf{t}_Q = \mathbf{centriod}(R_Q) = \frac{1}{|R_Q|}\sum_{j=1}^{|R_Q|}\mathbf{vector}(p_j). \quad (1)$$

Obviously, for each modeled site $s_Q^j \in \mathcal{B}_Q$, its topic $\mathbf{t}_j$ can be represented as follows:

$$\mathbf{t}_j = \mathbf{vector}(\mathbf{pseudo-page}(s_Q^j.C)) \quad (2)$$

where **pseudo-page**() denotes the operation that concatenates all pages in a site's page collection to form a pseudo-page.

Thus, the *relevance weight* of site s_Q^j (node j) can be computed by:

$$\mathbf{w}[j] = \mathbf{similarity}(\mathbf{t}_j, \mathbf{t}_Q) = \frac{\mathbf{t}_j \mathbf{t}_Q}{|\mathbf{t}_j| \times |\mathbf{t}_Q|}. \quad (3)$$

All the weights $\mathbf{w}[j]$ form a positive vector $\mathbf{w}$.

According to the above ideas and approaches, we have designed the following node-weighting algorithm **Weighting**: Table b

Algorithm **Weighting**($\mathcal{B}_Q$, R_Q)
//Algorithm for weighting nodes with content analysis.
$\mathcal{B}_Q = \{s_Q^1, s_Q^2, ..., s_Q^m\}$: a site base-set containing m sites.
R_Q: a page root-set.
1. Represent the topic of query Q as $\mathbf{t}_Q := \mathbf{centroid}(R_Q)$.
2. FOR $j := 1$ TO m DO
3. Represent the topic of site s_Q^j as $\mathbf{t}_j := \mathbf{vector}(\mathbf{pseudo\text{-}page}(s_Q^j.C))$.
4. FOR $j := 1$ TO m DO
5. Set relevance-weight $\mathbf{w}[j] := \mathbf{similarity}(\mathbf{t}_j, \mathbf{t}_Q)$.
6. RETURN $\mathbf{w}$.

3.4 Algorithm s-HITSc

The process of site-granularity topic distillation using our algorithm s-HITSc falls into two phases: First, *graph construction*, building a query-specific site neighborhood

graph and weighting all nodes on the graph; Second, *weighted iteration*, performing hyperlink analysis on the node-weighted graph to identify top authorities and hubs on the topic.

Using the above algorithms **Aggregation** and **Weighting**, the following algorithm **SGraph** can construct the node-weighted site graph.

To combine content analysis with hyperlink analysis in the iteration of our algorithm, the ***I*** and ***O*** operations used by original algorithm HITS must be modified into two *weighted operations*:

$\boldsymbol{I}_w$ operation: $\mathbf{a}[u] := \sum_{\{v|(u,v)\in E\}} (\mathbf{w}[v] \times \mathbf{h}[v])$, and
$\boldsymbol{O}_w$operation: $\mathbf{h}[u] := \sum_{\{v|(u,v)\in E\}} (\mathbf{w}[v] \times \mathbf{a}[v])$,

where $\mathbf{w}[v]$ is the *relevance weight* of node v. This gives us the following algorithm **WIteration** for hyperlink analysis controlled by the weights. Table c, Table d

Algorithm **SGraph**(Q, ε ,r, d)
//Algorithm for constructing a site neighborhood-graph and weighting the nodes.
Q: a broad-topic query.
ε: a text-based search engine or a text index system.
r, d: natural numbers.
1. Let root-set R_Q denote the top r results of ε on Q.
2. Initialize page base-set $B_Q := R_Q$.
3. For each page $p \in R_Q$, do:
4. [Add all pages hyperlinked to by p to B_Q.
5. Add at most d pages hyperlinking to p to B_Q.].
6. Form site base-set $\mathcal{B}_Q$:= **Aggregation**(B_Q).
7. Set weight vector **w** := **Weighting**($\mathcal{B}_Q$, R_Q).
8. Induce site neighborhood-graph G by $\mathcal{B}_Q$.
9. Return G.

Algorithm **WIteration**(G, k, c)
//Algorithm for reporting top-ranked authorities & hubs.
k, c: natural numbers.
1. Initialize vectors **a** and **h** to $\mathbf{z} = (1, 1, ..., 1)^T \in R^n$.
2. FOR i :=1 TO k DO
3. [Apply the $\boldsymbol{I}_w$ operation on **h** to compute **a**.
4. Apply the $\boldsymbol{O}_w$ operation on **a** to compute **h**.
5. Normalize **a** and **h** to unit vectors.].
6. Report the nodes with the c largest coordinates in vector **a** as *authorities*, and the nodes with the c largest coordinates in vector **h** as *hubs*, respectively.

4 Theoretical analysis

Theoretical analysis with linear algebra shows that the authority and hub vectors will eventually converge after arbitrarily large number of iterations in original algorithm HITS [13]. Although the underlying matrix computation of the iteration in our algorithm s-HITSc is slightly different from that of HITS, we can draw the following conclusion:

Theorem 1. The hub vector **h** and authority vector **a** will eventually converges when the iteration is performed with arbitrarily large values of k, furthermore, they converge to such limits that the values of their components are reasonably adjusted by the nodes' relevance-weights.

Proof. Let n be the size of the neighborhood graph $G = (V, E)$. The authority and hub (column) vector **a**, **h** $\in R^n$, and the relevance-weight vector $\mathbf{w} \in R^n$ ($0 < \mathbf{w}[i] \le 1$, $i = 1, 2, ..., n$). Construct matrix $\mathbf{X} = (x_{ij})_{n\times n}$ and $\mathbf{Y} = (y_{ji})_{n\times n}$ such that $x_{ij} = \mathbf{w}[j]$ and $y_{ji} = \mathbf{w}[i]$ iff $(i, j) \in E$, otherwise $x_{ij} = 0$ and $y_{ji} = 0$. Then step 3 and step 4 of algorithm **WIteration** can be rewritten as: $\mathbf{a} = \mathbf{Yh}$ and $\mathbf{h} = \mathbf{Xa}$, respectively. Let $\mathbf{z} = (1,1,...,1)^{\mathrm{T}} \in R^n$ (T denotes the transpose of a matrix or a vector), thus, after k iterations, **a** and **h** are the unit vectors in the directions of $(\mathbf{YX})^{k-1}\,\mathbf{Yz}$ and $(\mathbf{XY})^k\,\mathbf{z}$, respectively. Let $\mathbf{M} = \mathbf{XY} = (m_{ij})_{n\times n}$. Note that **M** is a non-negative matrix and its entries m_{ii} are not all zero, and $m_{ij} > 0$ iff $m_{ji} > 0$. Thus, **M** can be transformed into a block diagonal matrix **N** by some row-column permutations, i.e., there exists a permutation matrix **P** such that $\mathbf{PMP}^T = \mathbf{N} = \mathrm{diag}[\mathbf{N}_{11}, \mathbf{N}_{22}, ..., \mathbf{N}_{ll}, \mathbf{O}]$, where, each block $\mathbf{N}_{ii}$ ($i = 1, 2, ..., l$) is a positive matrix, and **O** is a zero matrix (i.e., null matrix).

According to Perron Theorem [7, 8], each positive matrix $\mathbf{N}_{ii}$ ($i = 1, 2, ..., l$) must have a unique real positive eigenvalue $\lambda_1^{(i)}$ of multiplicity 1, and it is rigidly greater than the moduli of other eigenvalues $\lambda_j^{(i)}$ ($j = 2, 3, ..., S_i$, where S_i is the order of $\mathbf{N}_{ii}$), and there must exist a positive eigenvector (i.e., the *principle eigenvector*) $\xi_1^{(i)}$ belonging to $\lambda_1^{(i)}$. Consider vector $\mathbf{N}^k\mathbf{z} = ((\mathbf{N}_{11}^k\mathbf{z}_1)^{\mathrm{T}}, (\mathbf{N}_{22}^k\mathbf{z}_2)^{\mathrm{T}}, \ldots, (\mathbf{N}_{ll}^k\mathbf{z}_l)^{\mathrm{T}}, \mathbf{O}^{\mathrm{T}})^{\mathrm{T}}$, where, $\mathbf{z}_i = (1, 1, \ldots, 1)^{\mathrm{T}} \in R^{S_i}$, **O** is a zero vector, k is the times of the iteration. Suppose that all the eigenvectors of $\mathbf{N}_{ii}$: $\xi_1^{(i)}, \xi_2^{(i)}, \ldots, \xi_{S_i}^{(i)}$ can form a base in R^{S_i} (otherwise, similar proof can be given via Jordan canonical form, here we leave out for saving space), then $\mathbf{z}_i = \sum_{j=1}^{S_i} c_j^{(i)}\xi_j^{(i)}$ Thus, $\mathbf{N}_{ii}^k\mathbf{z}_i = \sum_{j=1}^{S_i}(c_j^{(i)}\lambda_j^{(i)^k})\xi_j^{(i)}$, and thus vector $\mathbf{N}^k\mathbf{z} = (\sum_{j=1}^{S_1}(c_j^{(1)}\lambda_j^{(1)^k})\xi_j^{(1)^{\mathrm{T}}}, \sum_{j=1}^{S_2}(c_j^{(2)}\lambda_j^{(2)^k})\xi_j^{(2)^{\mathrm{T}}}, \ldots, \sum_{j=1}^{S_l}(c_j^{(l)}\lambda_j^{(l)^k})\xi_j^{(l)^{\mathrm{T}}}, \mathbf{O}^{\mathrm{T}})^{\mathrm{T}}$. Since each $\mathbf{N}_{ii}$ must have a greatest positive eigenvalue $\lambda_1^{(i)}$ of multiplicity 1 ($i = 1, 2, ..., l$), there must exist $\lambda_1^{(t)} = \max\{\lambda_1^{(i)}, 1 \le i \le l\}$, $1 \le t \le l$. Now consider vector $(1/\lambda_1^{(t)^k})\mathbf{N}^k\mathbf{z} = (\sum_{j=1}^{S_1}[c_j^{(1)}(\lambda_j^{(1)}/\lambda_1^{(t)})^k]\xi_j^{(1)^{\mathrm{T}}}, \sum_{j=1}^{S_2}[c_j^{(2)}(\lambda_j^{(2)}/\lambda_1^{(t)})^k]\xi_j^{(2)^{\mathrm{T}}}, \ldots, \sum_{j=1}^{S_l}[c_j^{(l)}(\lambda_j^{(l)}/\lambda_1^{(t)})^k]\xi_j^{(l)^{\mathrm{T}}}, \mathbf{O}^T)^T$. When k increase infinitely, its components $\sum_{j=1}^{S_i}[c_j^{(i)}(\lambda_j^{(i)}/\lambda_1^{(t)})^k]\xi_j^{(i)}$ converge either to **O** (for $i \ne t$ and $\lambda_1^{(i)} < \lambda_1^{(t)}$) or to $c_1^{(i)}\xi_1^{(i)}$ (for $\lambda_1^{(i)} = \lambda_1^{(t)}$). So vector

Table 1 Top ten authorities for query 'WTO' found by HITS (The graph modeled 6,154 pages. Date 03-10-2)

Rank	Rank in Root-set	Authorities page (URL)	Title of page	Type	Authority score
1	–	http://seattle.indymedia.org/	Seattle Indymedia	N	0.091
2	–	http://www.indymedia.org/	Independent Media Center | http://www.indymedia.org |(((i)))	N	0.080
3	–	http://cancun.mediosindependientes.org/	Medios Independientes Cancun: home	N	0.061
4	–	http://mexico.indymedia.org/	Indymedia México-Home	N	0.053
5	–	http://chiapas.mediosindependientes.org/	Centro de Medios Independientes, Chiapas	N	0.053
6	–	http://ontario.indymedia.org/	TWiki. Ontario. WebHome	N	0.042
7	–	http://enemycombatantradio.org/	enemy combatant radio | san francisco indymedia radio	N	0.041
8	–	http://tech.sfimc.net/	sf-active/indymedia project page	N	0.033
9	–	http://cancun.mediosindependientes.org /newswire/display/238/index.php	Medios Independientes Cancun: newswire/238	N	0.012
10	–	http://www.cancuncommittee.org/	Bienvenidos a Canc	N	0.012

$(1/\lambda_1^{(t)^k})\mathbf{N}^k\mathbf{z}$, thus $\mathbf{N}^k$ $\mathbf{z}$ converges. Since $\mathbf{h}$ is the unit vector in the direction of $(\mathbf{XY})^k\,\mathbf{z} = \mathbf{M}^k\,\mathbf{z} = \mathbf{P}^{\mathrm{T}}\,\mathbf{N}^k\,\mathbf{Pz} = \mathbf{P}^{\mathrm{T}}\,\mathbf{N}^k\,\mathbf{z}$, it follows that $\mathbf{h}$ converges. Similarly, vector $\mathbf{a}$ converges.

Now, we consider the limit to which $\mathbf{h}$ converges. For each block $\mathbf{N}_{ii} = (n_{uv})$, $(u, v = 1, 2, ..., S_i)$, we know that $\lambda_1^{(i)} = \rho(\mathbf{N}_{ii})$, where $\rho(\mathbf{N}_{ii})$ is the spectral radius of $\mathbf{N}_{ii}$. According to Frobenius Theorem [7, 8], we have $\min\{\sum_{v=1}^{S_i} n_{uv}, 1 \le u \le S_i\} \le \lambda_1^{(i)} \le \max\{\sum_{v=1}^{S_i} n_{uv},\ 1 \le u \le S_i\}$. Note that each $\sum_{v=1}^{S_i} n_{uv}$ thus $\lambda_1^{(i)}$ is related not only to the link density of the corresponding community on the graph but also to the relevance-weights of the nodes in the community, and $\lambda_1^{(i)}$ tends to be augmented by the higher weights. Thus, see the proof above, vector $\mathbf{h}$ converges to such a limit that the values of its components are reasonably adjusted by the nodes' relevance-weights. So does vector $\mathbf{a}$.

5 Experiments on the Web

We have developed a prototype system implementing our algorithm s-HITSc and carried out twice experiments in the context of the Web using the system. The first-time experiment (during 3∼7 January 2003) used query topics: (1) Abortion, (2) XML Schema, (3) Rock Climbing, (4) Stamp Collecting, (5) Gardening, which were mainly selected from [3, 4, 15]. The second experiment (during 2∼3 October 2003) used query topics: (1) WTO, (2) Genetic, (3) Severe Acute Respiratory Syndrome.

In the experiments, the root-set URLs and the URLs that hyperlink to the root-set pages were fetched from the search engine Fast Search (http://www.alltheweb.com). When adding neighboring-pages hyperlinked to by the root-set pages to the base-set (step 4 of **SGraph**), we eliminated such pages that they and the source pages are on the same sites, since the intrinsic hyperlinks often serve a navigational function [5, 13]. We set parameter $k = 50$ because we found the iteration converges quickly. Other parameters were set as follows: $r = 200$; $d = 50$; $c = 10$. More information about the experiments and the results can be found at the Web site of our experiments [2].

On the whole, the earlier and the later experiments show a similar result and comfirm the conclusions of theoretical analysis. Here, we compare the HITS algorithm and our improved s-HITSc algorithm to show the effectiveness of our algorithm based on the results of the later experiment. Tables 1∼3 and Tables 4∼6 list all the top ten authorities on the three query topics found by HITS and s-HITc, respectively. In the tables, column 'Rank in Root-set' shows the pages' ranks in the root-set (produced by Fast Search); column 'Type' shows the types of pages: pages in the root-set (denoted by 'R') and neighboring pages of the root-set pages (denoted by 'N').

All the problems mentioned early emerge in the results of HITS algorithm. *First*, it fails to meet users' site-granularity information needs. Among the top ten authorities on query 'WTO' in Table 1, the authorities ranked 3 and 9 are of the same Web site (http://cancun.ediosindependientes.org/). Similar situations can be found in Table 3. Those authorities that are of same sites can be aggregated into single authority-sites.

Second, HITS tends to produce unreasonable and irrelevant results. The authority ranked 2 on query topic 'WTO' in Table 1 is the homepage of a synthesized news site (http://www.indymedia.org/) which is not relevant to the query topic. Actually, there exists a high-scored hub (actually, scored 0.881) hyperlinking to this page but the meaning of the hyperlink is not about the topic 'WTO', which results in an unreasonable high authority-score being given to the page in HITS algorithm. Similar situation appears on the other two queries. The authority ranked 8 on query 'Genetic' in Table 2 is an obviously unmeaningful and unreasonable result—it is an Acrobat Reader download page. The autority ranked 2 on query 'Severe Acute Respiratory Syndrome' in

[2] http://cse.seu.edu.cn/labs/dbgroup/resource/TD-experiments/main.htm

Table 2 Top ten authorities for query 'Genetic' found by HITS (The graph modeled 9,359 pages. Date 03-10-3)

Rank	Rank in root-set	Authorities page (URL)	Title of page	Type	Authority score
1	66	http://www.cgdn.generes.ca/	yanadian genetic diseases network	R	0.168
2	–	http://www.hum-molgen.de/	international communication forum in human molecular genetics	N	0.154
3	–	http://www.jax.org/	The Jackson Laboratory. Advance research on human health	N	0.144
4	–	http://www.ncgr.org/	National Center for Genome Resources	N	0.144
5	–	http://www.ncbi.nlm.nih.gov/	NCBI:a national resource for molecular biology information	N	0.131
6	–	http://www.er.doe.gov/	US department of energy	N	0.127
7	–	http://www.aaas.org/	American association for the advanced of science	N	0.117
8	–	http://www.adobe.com/products /acrobat/readstep2.html	Adobe Acrobat Reader – Download	N	0.114
9	–	http://doegenomes.org	gonome programs of the U.S. Department of Energy Office of Science	N	0.113
10	–	http://doegenomestolife.org	Genomes to Life program,US Department of Energy	N	0.101

Table 3 is the content-table Web page of a medical textbook which is also not quite relevant to the query topic.

Last, to a multiple-topic query or a broad-topic query, HITS tends to find those authorities which focus on one of the topics or one aspect of the broad topic. Among the top ten authorities on the query topic 'WTO' (a multiple-topic query), the authorities ranked from 1 to 6 are all pages about WTO trade news. It is because there are abundant links between high-scored hubs and these authority pages, which leads to occurring a TKC effect. Similarly, the top ten authorities on query 'Severe Acute Respiratory Syndrome' in Table 3 are all pages of global official site focusing on SARS control, not pertaining to other aspects of the query topic.

In contrast, the authorities on the three queries found by our improved algorithm s-HITc show much better results than HITS. *First*, see Tables 4∼6, after aggregating to site granularity, there are many pages being aggregated into one authority-site, which can meet the users' site-granularity information needs.

Second, all authority and hub sites on the three topics found by s-HITSc were browsed and the first-rate quality of these sites was confirmed by the topic experts from our two universities. The judgement indicates that s-HITSc has produced relevant and meaningful results.

Last, the authorities found by s-HITSc include various topical aspects of a given query. See Table 4, the top ten authorities on query 'WTO' can be classified into two topics: *World Trade Organization* site (ranked 1, 2,

Table 3 Top ten authorities for query 'Severe Acute Respiratory Syndrome' found by HITS (The graph modeled 4,787 pages. Date 03-10-3)

Rank	Rank in root-set	Authorities page (URL)	Title of page	Type	Authority score
1	98	http://www.who.int/csr/sars/guidelines/en/	WHO-SARS	R	0.050
2	–	http://www.emedicine.com/med/ contents.htm	eMedicine Medical Textbooks - Medicine, Ob/Gyn, Psychiatry, and Surgery - Free Physician Reference Articles and Texts	N	0.049
3	–	http://www.emedicine.com/med/ topic3662.htm	eMedicine - Severe Acute Respiratory Syndrome (SARS): Article by Richard L Oehler, MD, FACP	N	0.047
4	–	http://www.cdc.gov/	CDC Centers for Disease Control and Prevention Home Page	N	0.046
5	–	http://www.cdc.gov/ncidod/sars/ clinicians.htm	CDC | Clinicians/Healthcare Settings-Severe Acute Respiratory Syndrome (SARS)	N	0.045
6	–	http://www.cdc.gov/ncidod/sars/ casedefinition.htm	CDC | SARS Case Defintion - Severe Acute Respiratory Syndrome	N	0.045
7	–	http://www.cdc.gov/ncidod/sars/ diagnosis.htm	CDC | Clinical Evaluation and Diagnosis-Severe Acute Respiratory Syndrome (SARS)	N	0.045
8	–	http://www.cdc.gov/ncidod/sars/ specimen_5fcollection_5fsars2.htm	CDC | specimen_5fcollection_5f - Severe Acute Respiratory Syndrome (SARS)	N	0.045
9	–	http://www.cdc.gov/ncidod/sars/ infectioncontrol.htm	CDC | infection control-Severe Acute Respiratory Syndrome (SARS)	N	0.045
10	–	http://www.cdc.gov/ncidod/sars/ic.htm	CDC | Infection Control-Severe Acute Respiratory Syndrome	N	0.030

Table 4 Top ten authorities on the topic 'WTO' found by s-HITSc (The graph modeled 3,616 sites aggregated from 6,154 pages. Date 03-10-3)

Rank	Selected identifying page of the site				Num. of aggregated pages	Relevance weight	Authority score
	URL	Title	Type	Rank in root-set			
1	http://www.wto.org/	WTO | Welcome to the WTO website	R	1	84	0.717	0.350
2	http://santacruz.indymedia.org/	santa cruz indymedia: home	N	–	55	0.708	0.293
3	http://www.intracen.org/	ITC: the technical cooperation agency of UNCTAD and WTO for operational, enterprise-oriented aspects of trade development	R	5	14	0.671	0.280
4	http://www.worldtradelaw.net/	WorldTradeLaw.net-The Online Source for World Trade Law	R	53	44	0.416	0.260
5	http://www.world-tourism.org/	World Tourism Organization (WTO)	R	2	19	0.612	0.259
6	http://www.llrx.com/	Law and technology news on WTO	N	–	8	0.714	0.257
7	http://www.itd.org/	Trade& Development Centre: A Joint Venture of the World Bank and the World Trade Organization	R	10	3	0.657	0.203
8	http://www.usitc.gov/taffairs.htm	United States International TRade Commission	R	59	11	0.675	0.200
9	http://www.oecd.org/ech/	Organisation for Economic Co-operation and Development	N	–	4	0.651	0.184
10.	http://www.wtowatch.org/	IATP | Trade Observatory	R	3	32	0.753	0.183

Table 5 Top ten authorities on the topic 'Genetic' found by s-HITSc (The graph modeled 6,422 sites aggregated from 9,359 pages. Date 03-10-3)

Rank	Selected identifying page of the site				Num. of aggregated pages	Relevance weight	Authority score
	URL	Title	Type	Rank in root-set			
1	http://www.kumc.edu/gec/geneinfo.html	Information for genetic professionals, University of Kansas Medical Center	R	98	36	0.651	0.347
2	http://www.ncbi.nlm.nih.gov/	NCBI: a national resource for molecular biology information	N	–	35	0.769	0.172
3	http://www.geneclinics.org/	a publicly funded medical genetics information resource developed for physicians	R	39	8	0.527	0.160
4	http://www.geneticalliance.org/	The definitive resource for reliable genetics information	R	1	23	0.697	0.103
5	http://www.turner-syndrome-us.org/	turner syndrome: a genetic related disease	R	87	3	0.536	0.096
6	http://www.chromodisorder.org/	a non-profit organization providing information& support for Families and Professionals affected by Chromosome Deletions, Trisomies, Inversions, Translocations and Rings.	R	127	8	0.699	0.082
7	http://www.icomm.ca/geneinfo/res.htm	a web site provide genetic info	N	–	15	0.589	0.028
8	http://www.mostgene.org/	The MoSt GeNe website is for sharing medical genetics knowledge and resources with other health care practitioners, patients, and caregivers.	R	12	9	0.546	0.021
9	http://www.faseb.org/genetics/	A World of Genetics Societies	R	59	8	0.593	0.014
10	http://www.aaksis.org/	American Association for Klinefelter Syndrome Information and Support	R	168	7	0.524	0.010

Table 6 Top ten authorities on the topic 'Severe Acute Respiratory Syndrome' found by s-HITSc (The graph modeled 2,322 sites aggregated from 4,787 pages. Date 03-10-2)

Rank	Selected identifying page of the site				Num. of aggregated pages	Relevance weight	Authority score
	URL	Title	Type	Rank in root-set			
1	http://www.who.int	Homepage (World Health Organization)	N	–	103	0.845	0.245
2	http://www.cdc.gov/	U.S. Centers for Disease Control& Prevention.	N	–	112	0.844	0.215
3	http://www.hc–sc.gc.ca/	Health Canada	N	–	54	0.861	0.123
4	http://www.fda.gov/	US Food and drug admission	N	–	5	0.777	0.117
5	http://www.info.gov.hk/eindex.htm	Hong Kong SAR Government Information Centre	N	–	8	0.609	0.110
6	http://www.doh.gov.uk/	Homepage (Department of Health of United Kingdom)	N	–	6	0.865	0.107
7	http://content.nejm.org	The New England Journal of Medicine (NEJM) is a weekly general medical journal	N	–	30	0.673	0.093
8	http://www.health.gov.au/sars.htm	Australian Government Department of Health and Ageing - SARS - Severe Acute Respiratory Syndrome	R	6	21	0.569	0.085
9	http://www.city.toronto.on.ca/health/sars/	SARS (Severe Acute Respiratory Syndrome) - Toronto Public Health	N	–	5	0.818	0.084
10	http://www.nlm.nih.gov/medlineplus/severeacuterespiratorysyndrome.html	MEDLINEplus: Severe Acute Respiratory Syndrome	R	4	119	0.851	0.070

3, 4, 6, 7, 8, 9, 10) and *World Tourism Organization* site (ranked 5). And on the topic *World Trade Organization*, we can further classify them into three sub-topics: *WTO news* (ranked 2, 10, 6), *WTO trade development* (ranked 1, 3, 7, 9), and *WTO laws* (ranked 4, 6, 8). This indicates that these authorities pertain not only different topics on a multiple-topic query but various aspects of a topic, which shows that our algorithm trends to more "balanced". Other two queries have similar results. See Table 6, the top ten authorities on query 'Severe Acute Respiratory Syndrome' contain: (1) Three global official site related to the query topic (ranked 1, 2, 3): *World Health Organization (WHO)*, *Center of Disease Control (CDC)* and *Health Canada*; (2) Three national official site (ranked 4, 6, 8): *US Food and Drug Admission*, *Department of Health of United Kingdom*, and *Australian Government Department of Health and Ageing*; (3) A Web portal from Hong Kong related to SARS (ranked 5); (4) Another three news and publication sites about SARS (ranked 7, 9, 10). Similarly, see Table 5, the top ten authorities on query 'Genetic' can be classified into three sub-topics: (1) genetics (ranked 8, 9); (2) genetic disease (ranked 5, 6, 10); (3) genetic information (ranked 1, 2, 3, 4, 7).

In conclusion, the experimental results show that: (1) s-HITSc is effective to identify meaningful and quality authority and hub sites on a given topic; (2) Unlike more "focused" algorithms (such as HITS) that are easy to result in topic drift especially to multiple-topic queries or broad-topic queries [4, 15], s-HITSc trends to be more "balanced" and identify the authorities pertain to various topical aspects of a given query; (3) During analysis, large quantity of pages are aggregated into sites by s-HITSc, which defines equal influence weights for all authors of information sources (Web sites) and meets the users' site-granularity information needs.

6 Conclusions

In addition to viewing the Web as a "page graph", which tends to define unjust influence weights for different authors of information sources and fails to meet users' site-granularity information needs, traditional topic distillation algorithms such as HITS are cocksure of the two assumptions (*quality* and *relevance* presumptions) and use "link-only" methods, which lead to producing unreasonable and irrelevant results. Based on the analysis of three deficiencies of classical topic distillation algorithm HITS (*failing to meet users' site-granularity information needs, tending to produce unreasonable results, and topic drift*), this paper presents an improved model and algorithm s-HITSc. Theoretical analysis and experimental results show that the new algorithm can control topic drift and identify more reasonable and meaningful authority and hub sites on a given query topic.

In practice, s-HITSc algorithm can be applied to the following scenarios: (1) As an optional *post-retrieval* process, the algorithm can be integrated into crawler-based search engines to provide users with site-granularity topic distillation services; (2) As a *pre-retrieval* process, the algorithm can provide (semi)automatic finding and indexing of authoritative topical sites for human-powered directories or vertical portals; (3) The algorithm can provide inputs (topical sites and related

pages) for *task-based site search* systems (a prototype system, see [16]) envisioned by Hearst [9] as next generation Web search services.

Although the combination of content analysis and hyperlink analysis can provide better search services, these approaches have intrinsic limitations due to lacking of formal semantic descriptions for content and links on the Web. To achieve more elaborate, precise automated searches [18], the current Web must be evolved into the next generation—the Semantic Web.

Acknowledgements An earlier short version of this paper, "Site-granularity Topic Distillation on the Web by Combining Content and Hyperlink Analysis" appeared in *Proceedings of the Second International Conference on Machine Learning and Cybernetics*, IEEE Press, November 2003. We would like to thank the program committee of the conference for granting an Outstanding Paper Award to the paper and selecting this extended version of the paper to be published in the international journal *Soft Computing*. This work was partially funded by the Tenth Five-year High-tech Projects of Jiangsu Province of China under Grant No. BG2001013, and by the Natural Science Foundation of Jiangsu Province of China under Grant No. BK2003001. The findings and views expressed in this paper are those of the authors, and not necessarily of the funding organizations. The authors would like to thank Prof. *Daoyuan Zhu* (Department of Applied Mathematics, Southeast University) for his help in the proof of Theorem 1 in the paper, Prof. *Yuzhong QU* (Department of Computer Science and Engineering, Southeast University) for his helpful suggestions during the earlier research work, and all the topic experts from two universities for their arduous work.

References

1. Amento B, Terveen L, Hill W (2000) Does "authority" mean quality? Predicting expert quality ratings of Web documents. In: Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval, ACM Press, Athens, pp. 296–303
2. Baeza-Yates R, Ribeiro-Neto B (1999) Modern information retrieval. Addison Wesley, New York
3. Bharat K, Henzinger M (1998) Improved algorithms for topic distillation in a hyperlinked environment. In: Proceedings of the 21th annual international ACM SIGIR conference on research and development in information retrieval, ACM Press, Melbourne, pp. 104–111
4. Borodin A, Roberts GO, Rosenthal JS, Tsaparas P (2001) Finding authorities and hubs from link structure on the World Wide Web. In: Proceedings of the 10th international World Wide Web conference, ACM Press, Hong Kang, pp. 415–429
5. Chakrabarti S, Dom Byron E, et al (1999) Mining the Web's link structure. IEEE Computer 32(8):60–67
6. Craswell N, Hawking D, Robertson S (2001) Effective site finding using link anchor information. In: Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval, ACM Press, New Orleans, pp. 250–257
7. Golub G, Van Loan CF (1989) Matrix computations. Johns Hopkins University Press, Baltimore
8. Gong-Ning Chen (1990) Theory of matrix and its application. China Higher Education Press, Beijing
9. Hearst MA (2000) Next generation Web search: setting our sites. Bulletin of the technical Committee on data engineering, IEEE Computer Society 23(3):38–48
10. Henzinger M (2001) Hyperlink analysis for the Web. IEEE Internet Comput. 5(1):45–50
11. Jain AK, Murty MN, Flynn PJ (1999) Data clustering: A review. ACM Comput. Surv. 31(3):264–323
12. Jansen BJ, Pooch U (2001) A Review of Web searching studies and a framework for future research. J. Am. Soc. Inf. Sci. Technol. 52(3):235–246
13. Kleinberg J (1999) Authoritative sources in hyperlinked environment. J. ACM 46(5):604–632
14. Kleinberg J, Lawrence S (2001) The structure of the Web. Science 294(30):1849–1850
15. Lempei R, Moran S (2001) SALSA - the stochastic approach for link-structure analysis. ACM Trans. Inf. Sys. 19(2):131–160
16. Levene M, Wheeldon R (2001) A Web site navigation engine. In: Poster Proceedings of the 10th international World Wide Web conference, ACM Press, Hong Kong, pp. 1014–1015
17. Search Engine Watch. http://www.searchenginewatch.com
18. Urvi Shah, Timothy W Finin, Anupam Joshi (2002) Information retrieval on the semantic web. In: Proceedings of the ACM 11th international conference on information and knowledge management, ACM Press, McLean, Virginia, USA, pp. 461–468