行政院國家科學委員會專題研究計畫 成果報告

運用於網際網路上企業環境適性化偵搜之遞增式資訊興趣探勘

計畫類別： 個別型計畫

計畫編號： NSC 92－2213－E－216－018

執行期間： 92 年 08 月 01 日 至 93 年 07 月 31 日

執行單位： 中華大學資訊管理學系

計畫主持人： 劉瑞瓏

計畫參與人： 楊怡雰, 劉玟杰

報告類型： 精簡報告

處理方式： 本計畫可公開查詢

中 華 民 國 93 年 10 月 23 日

行政院國家科學委員會專題研究計畫成果報告

運用於網際網路上企業環境適性化偵搜之遞增式資訊興趣探勘

Incremental Mining of Information Interest for Adaptive Scanning of Business Environments through the Internet

計畫編號：NSC 92-2213-E-216-018
執行期限：92 年 8 月 1 日至 93 年 7 月 31 日

主持人：劉瑞瓏　中華大學資訊管理學系

1. 中文摘要

企業常將其感興趣的企業內外資訊以階層式資料匣的方式分門別類來儲存。此專屬於個別企業之階層式資料匣不但方便了資訊之管理與運用，並同時反映了該企業之中長期資訊興趣。一個資料匣可對應一種資訊興趣型態。針對每種興趣型態，企業需要持續地從網際網路上偵搜環境中與此興趣型態相關之資訊異動，以隨時充分掌握契機並據以作出適當應變。為達此目的，每個興趣型態均需有明確且可讀性高之興趣描述，才能隨時導引系統善用有限之資源(如網路頻寬及伺服器負載等)來發掘並監控相關資訊。然而企業往往無法為每一興趣型態明確地定義其描述。事實上，一個興趣型態的描述是整合隱藏於對應資料匣內眾多的文件中。更有甚者，此興趣描述仍會因這些文件之異動而變遷。因此，本研究計劃在我們以往文件探勘、使用者需求釐定、與環境偵搜與的研究基礎上，進一步研發一個遞增式文件探勘技術，自動探勘出各興趣型態之描述，讓系統能依企業最新之興趣而適性調整其環境偵搜的策略。每一興趣型態之描述均需明確地表明其場景(背景)，並以可讀的方式表達，以利使用者了解與修正。此描述並可直接由一般搜尋引擎所接受，以期能廣泛地獲得偵搜之適當起始網站，提昇偵搜之成本效益。我們並在網際網路上實際驗證此技術。本研究成果可免除企業在明確描述興趣與廣泛提供起始網站上的困難，並讓企業源源不斷地獲得來自環境之相關資訊，掌握與企業息息相關之環境異動。

關鍵詞：資訊興趣之場景、明確之興趣描述、可讀之興趣描述、遞增式文字探勘、適性化環境資訊偵搜

Abstract

Businesses often hierarchically organize their internal and environmental information of interest (IOI) into folders (or categories). Such personalized hierarchical folders may not only facilitate the management of IOI, but also reflect the interest of each individual business. A folder corresponds to an interest type. The interest is relatively long-term when compared with one-shot queries. For such interest, environmental scanning through the Internet (ESI) should be a continuous job directed by the specifications of the interest. The specifications should be both *precise* and *comprehensible* in order to make ESI more cost-effective and controllable. However, expressing such specifications are quite difficult for the business, since each interest type is implicitly and collectively defined by the content (i.e. documents) of the corresponding folder, which may also evolve over time. In this project, based on our previous experiences in information need identification, text mining, and ESI, we develop an *incremental* text mining technique to identify the business's current interest by mining the business's information folders, making ESI more *adaptive* to the business's evolving interest. The specification mined for each interest type specifies the *context* of the interest type in suitable form (e.g. *conjunction-of-disjunctions* form), which is easy for business users to comprehend and refine. It helps the scanner to comprehensively start from proper seed sites and focus on those sites that are more likely to provide the information really of the business's interest.

The business may thus maintain her folders to constantly get IOI without paying much attention to the difficult tasks of interest specification and seed identification.

2. Introduction

Environmental scanning is a fundamental task for many business activities (e.g. decision making and product development) in order to keep the business competitive in the ever-changing world. Since much environmental information has been published on the Internet, environmental scanning through the Internet (ESI) has been a major way for a business to collect environmental information. It aims to *continuously* scan for information of interest (IOI) on the web information space. The IOI is relatively long-term (when compared with one-shot queries), since the jobs and preferences of a business are relatively long-term. Each business constantly requires timely information concerning her jobs and preferences. Given a specification of the business's interest, the scanner needs to identify the seeds from which IOI gathering and monitoring are started and conducted as a routine job. The interest may evolve when internal conditions (e.g. business strategy and preferences) and external conditions (e.g. the environments) change. The information scanned, together with the information manually collected, is often stored into folders (or categories) that are organized hierarchically. The personalized hierarchical folders facilitate the browsing and retrieval of the information.

Proper specification of the user's interest is obviously a key to ESI. It should be both *comprehensible* and *precise*. A comprehensible specification may facilitate manual refinement and management, while a precise specification may direct the scanner's effort to those spaces that deserve scanning. Imprecise specifications may significantly deteriorate the performances of most information gatherers on the web. Moreover, as the business's interest evolves, the specification may evolve as well. Improper or obsolete interest specification may consume lots of resources (e.g. efforts of the scanner's, bandwidths of the network, and services provided by the related information servers), while produce lots of garbage information to the business.

Unfortunately, the business often has difficulties in specifying her relatively long-term but evolving interest. The interest is actually implicitly defined in the hierarchy of information folders. A folder corresponds to an interest type, which is collectively defined by the documents under the folder (i.e. including the folder and subfolders of the folder). The business thus requires a mechanism that may identify and represent each interest type and its evolution. With the help provided by the mechanism, the business only needs to maintain her own folders. Proper specification of the business's interest may be automatically extracted to guide the scanner to find IOI using a smaller amount of resources.

Figure 1 illustrates the idea. Each business may set up and maintain a hierarchy of information folders. She may enrich the folders by new IOI that is either automatically scanned or manually collected. A folder thus corresponds to an interest type of the business. A set of folders may be designated (by the business) as an interest set for the scanner to work on. An interest miner is activated once the contents of the designated folders are updated. It identifies the newest specification of each interest type (folder). Based on the specification, the scanner identifies proper seeds to start continuous gathering and monitoring of IOI from the web. Therefore, ESI may thus be *adaptive* in the sense that scanning is actually adapted to the evolving interest of the business. With the recent successful development in the technology of information monitoring and gathering, the major challenge to realize the scanning flow lies on the interest miner, which is the core to achieve satisfactory and cost-effective ESI. This is the target explored in this project.

Based on our previous studies in information need identification, incremental text mining, and business environmental scanning, we aim to tackle the challenge by developing an incremental interest miner IMind (Interest Mining from personalized folders). IMind employs incremental text mining to derive *precise* and *comprehensible*

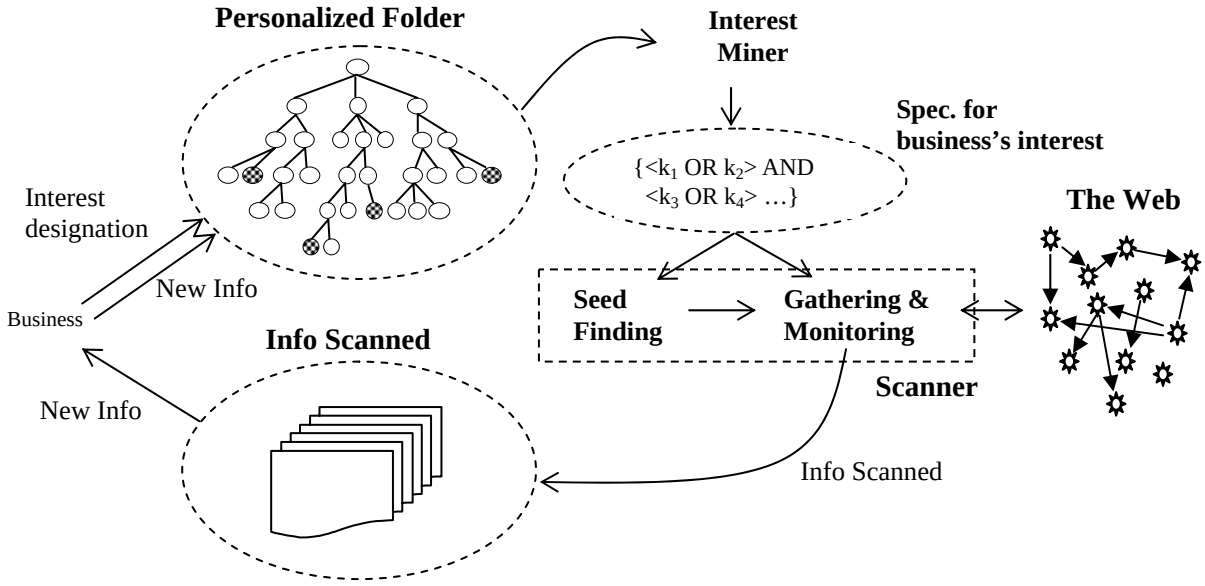

Figure 1. Mining personalized information folders for ESI

specifications from the business's personalized hierarchical information folders.

3. Result and Discussion

IMind was developed and evaluated in a real-world environment. It is to be published in [1]. In the research, IMind was shown to be efficient in deriving interest specifications to direct the scanner's effort to continuously finding the information that is really of the user's interest, without asking the user to conduct tedious (and even implausible) tasks of precise interest specification and seed identification.

IMind operates on a given hierarchy T of information folders. A folder corresponds to an interest type of the business. Each folder has a profile, which is initially empty. The business may designate a set G of interest types as the goals of ESI. Once a document p is added to a folder f, IMind triggers two functions (1) updating the profiles of related folders, and (2) notifying the scanner if the interest specification of any interest type in G is changed.

3.1 Profile mining

The profile of a folder f is a set of 3-tuples $<w, r_{w,f}, d_{w,f}>$, where w is a word (or feature), $r_{w,f}$ is the degree to which w may represent the content of the documents under f, and $d_{w,f}$ is the capability of w in discriminating f from siblings of f. A *good* profile term should be the one that is both *representative* (having a higher r-value) and *discriminative* (having a higher d-value).

Since $d_{w,f}$ is estimated with respect to *siblings* of f (rather than *all* folders as in most previous approaches), general (specific) terms tend to be good profile terms for higher- (lower-) level folders. For example, suppose in the business's hierarchy of folders, System Development (SD) has two subfolders: Decision Support Systems (DSS) and Accounting Information Systems (AIS). The terms like "information" and "maintenance" may have a lower d-value in both DSS and AIS (since the documents in both subfolders are about the implementation and maintenance of information systems), but a higher d-value in SD, if sibling folders of SD are not about system development. Therefore, "information" and "maintenance" may be good in discriminating SD from its siblings, which share the same generality levels. Similarly, the terms like "inventory" and "sales" may get a higher d-value in AIS, but not in SD or DSS. That is, based on the personalized hierarchy provided by the business, generality of each term is implicitly defined. IMind makes the generality explicit by mining.

Technically, the degree of representation of a word w under f (i.e. $r_{w,f}$) is estimated by $P(w|f)$, which is equal to $TF(w,f)$ / $Size(f)$, where $TF(w,f)$ is the times of occurrences of w in documents under f, and $Size(f)$ is the total number of terms in the documents under f (from the viewpoint of data mining, $P(w|f)$ may be viewed as the *support* of w under f). On the other hand, the capability of w in discriminating f from its siblings (i.e. $d_{w,f}$) is estimated by $P(w|f) * (B_f / \Sigma_i P(w|f_i))$, where B_f

is the number of siblings of f plus one (i.e. including f). The summation of $P(w|f_i)$ is conducted over f and its siblings. Thus, for example, w may get a higher d-value if it occurs frequently under f (i.e. $P(w|f)$ is high), but infrequently (on average) under siblings of f (i.e. $B_f / \Sigma_i P(w|f_i)$ is high). In that case, w may be good in discriminating f from siblings of f. Also note that, $0 \leq d_{w,f} \leq B_f$, for each profile term w in f. When w only occurs in f, $d_{w,f} = B_f$, while $d_{w,b} = 0$ for each sibling b of f. If $P(w|f)$ is higher than $\Sigma_i P(w|f_i)/B_f$ (i.e. the average $P(w|f_i)$), $d_{w,f} > 1$; otherwise $d_{w,f} \leq 1$.

3.2 Deriving interest specification

Profile mining provides a fundamental basis for deriving the specification of each folder (interest type). As noted above, it identifies good profile terms (i.e. both representative and discriminative) for each folder. General (specific) terms tend to have higher d-values in general (specific) folders. Therefore, good profile terms for ancestor folders of f may indicate the *context* of f. As an example, consider the above-mentioned folder hierarchy in which System Development (SD) has two subfolders: Decision Support Systems (DSS) and Accounting Information Systems (AIS). Through profile mining, "information" and "maintenance" may be found to be a good profile term for SD, while "sales" and "inventory" may be good profile terms for AIS. Thus a good specification for the AIS folder should include not only "sales" and "inventory," but also "information" and "maintenance," which may indicate the context of the folder.

Therefore, the specification for a folder f should be composed of good profile terms of f and each ancestor folder of f. The profile terms from the folders of the same pedigree should be integrated in conjunction form so that the requirement of each level of context generality may be enforced. On the other hand, the profile terms from each folder should be integrated in disjunction form so that the coverage of the terms may represent the main content of the folder. That is, the specification for a folder should be in *conjunction-of-disjunctions* form. For the above example, the specification for AIS may be like {… <information OR maintenance> AND <sales OR inventory>}.

IMind thus checks each goal of scanning and identifies those whose specifications are changed due to the newly added document. New specifications are derived and sent to the scanner immediately. For each goal folder f, the specification is derived by checking the profiles of f and ancestors of f. Good profile terms are extracted by selecting those terms that have higher r-values and d-values. The number of terms to be selected from a profile is governed by a system parameter α. The terms selected are then integrated in disjunction form. The final interest specification is simply the conjunction of the disjunctions for f and its ancestors. Note that, when selecting terms to form the disjunction for an ancestor of f, only the terms that occur in the profile of f may be the candidates. This method guarantees that each term in the specification derived for f really occurs in documents of f.

3.3 Experimental evaluation

We also evaluated IMind on a real-world text hierarchy and a search engine. The experiment aimed to evaluate the qualities of the interest specifications mined by IMind and several previous techniques. The qualities were measured by analyzing the seeds that may be retrieved by sending the interest specifications to the search engine. Interest specifications may be said to be more precise, if they may direct the search engine to find more seeds that are really of each interest type. That is, we employed a popular and well-developed search engine to objectively judge the qualities of the specifications produced by IMind and the previous techniques.

For objective evaluation and cross-validation, experimental data was extracted from the text hierarchy of Yahoo (http://www.yahoo.com). Without loss of generality, we extracted parts of the documents under several top-level categories. A category in the hierarchy corresponds to a folder. Based on the structure of the hierarchy, we constructed two hierarchies to serve as two businesses' hierarchies of information folders. This setting helped to investigate the performance of IMind under different hierarchies with different sizes, structures, and

contents. A document corresponds to a web site. It is a web page *manually* extracted to represent the main content of its corresponding web site (a web site often contains many pages not conveying its main content). The profile of each folder was built by mining the documents in the hierarchies. The two hierarchies were mined and tested independently. To conduct thorough performance investigation, all leaf folders were designated as the business's goals of scanning (i.e. the set *G* in Table 1).

For each folder designated as a goal, a specification was constructed by IMind and several previous techniques, and then sent to Yahoo to retrieve web sites. For each folder, we controlled the number (e.g. less than 200) of top-ranking web sites retrieved. For each site in the listing retrieved, Yahoo showed its brief summary, followed by its category (folder) in the whole hierarchy of Yahoo. Based on the information provided by Yahoo, we may evaluate the qualities of the specifications produced by IMind and the previous techniques.

We employed two evaluation criteria to measure the quality of the specifications: (1) the average number of web sites correctly retrieved per folder, and (2) the percentage of folders (in *G*) for which at least one web site is correctly retrieved. The first criterion aimed to measure the *completeness* of the retrieval of the information that is really of the individual goal interest types, while the second criterion aimed to measure the *reliability* of the retrieval across different interest types. Specifications for the folders may be said to be more precise, if they may direct the search engine to (1) retrieve more sites that are of individual goal folders, and (2) find sites for a higher percentage of goal folders.

The systems evaluated in the experiment included IMind and several baselines. IMind has only one parameter (i.e. α, the number of terms selected in each level of profile), which had two settings: 10 (IMind-10) and 20 (IMind-20). In addition to IMind, there were four baselines for performance comparison: Norm of the Folder (NOF), Rocchio (RO), Naive Bayes (NB), and Hierarchical Shrinkage (HS). For each leaf folder, the baselines created a profile by their individual weighting methods, although the profiles were originally used for other purposes (e.g. classification, filtering, and relevance feedback for query refinement).

It should be noted that, instead of incrementally maintaining an evolving feature set to express each folder's profile (as in IMind), all the baselines need to preset a fixed feature set on which each folder's profile is represented. The selection of the features was based on the strength of each feature, which was estimated by the χ^2 (chi-square) weighting technique. The technique has been shown to be more promising than several others. As noted in previous studies, the size of the feature set is an experimental issue (no standard way to set a perfect size). Therefore, for each baseline, we tested several different sizes. Also note that, since the profiles constructed by the baselines could not be comprehensible for the search engine, the baselines need to transform each profile into a query that is acceptable for the search engine. This was achieved by selecting and integrating those features having higher positive weights, since these features were more representative and discriminative than others. As in previous related studies, the features were integrated in a disjunction manner (i.e. the selected featured are integrated using OR). Moreover, since there is no standard way to determine how many features to select, to facilitate objective performance comparison with IMind, we controlled the length (i.e. number of terms) of the queries from the baselines. That is, as IMind, each baseline has two versions as well. For example, NOF-10 and NOF-20 were allowed to construct those queries having the same maximum lengths as those constructed by IMind-10 and IMind-20, respectively. Thus, for example, for a level-5 folder, the query constructed by NOF-10 may have the maximum length of 50 (= 5*10).

The experimental results showed that IMind is efficient, and it is capable of directing the scanner's effort to continuously finding the information that is really of the user's interest, without asking the user to conduct tedious (and even implausible) tasks of precise interest specification and seed identification.

4. Evaluation

ESI is triggered by the interest of each individual business. The interest is relatively long-term and evolving. Since the Internet information space is intrinsically huge and ever changing, a scanner for ESI often strives to consume lots of resources to continuously gather and monitor information of the user's interest. *Precise* and *comprehensible* specifications of the interest are thus essential from the perspectives of user satisfaction and cost-effectiveness of scanning. Unfortunately, the user often cannot express such specifications. The interest is actually *implicitly* and *collectively* defined by the *evolving* folders that contain those documents that deserve storing and referencing (e.g. documents related to the user's job descriptions, skills, and preferences). For each interest type, the business also has the difficulties in comprehensively identifying proper seeds to start scanning, since the web is intrinsically huge and ever changing. Improper and/or incomprehensive identification of seeds may significantly deteriorate the cost-effectiveness of ESI.

In this project, we successfully developed IMind, which is to be published in [1]. IMind guides ESI by incremental interest mining. The specification mined for each interest type is represented in conjunction-of-disjunctions form, which may facilitate comprehension and seed identification. Analyses and empirical results showed that the specifications mined by IMind may be more precise in expressing the context of each interest type. Through incremental mining of the specifications, IMind may also adapt the scanner to the evolving interest of the business. The delivery of such specifications to a scanner may keep the scanner precisely aware of the most recent interest of the business, directing each resource consumption and effort of the scanner to more focused and suitable targets, while separating the business from the tedious (even implausible) tasks of interest specification and seed identification.

Moreover, the development of IMind is actually an interdisciplinary study of text mining and business environmental scanning. From this point of view, the contributions of the project also include (1) mutual impacts to both text mining and business environmental scanning, (2) novel integration of the considerations from the above disciplines to develop an incremental mining technique for business environmental scanning, which is both essential and significant for business administration, and (3) training graduate and undergraduate students to integrate, design, implement, and apply what they have learned in order to practically satisfy the needs of businesses.

5. Reference

[1] Rey-Long Liu and Wan-Jung Lin, "Incremental Mining of Information Interest for Personalized Web Scanning," to appear in *Information Systems.*