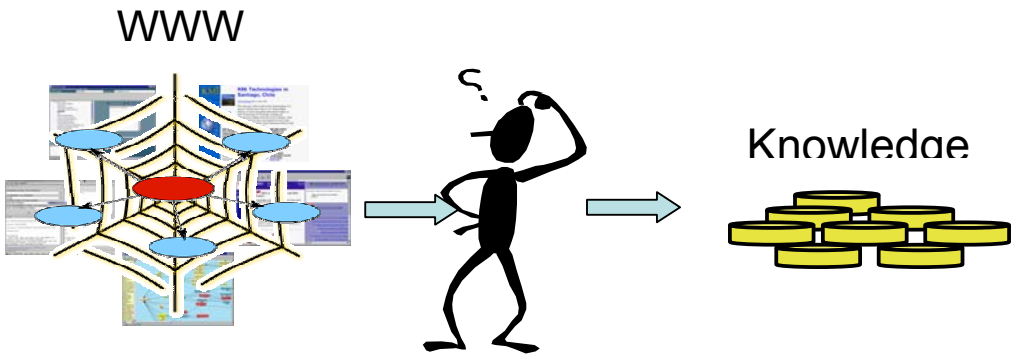
Mục lục

Chương 1. Tổng quan về khai phá dữ liệu Web và máy tìm kiếm

1.1. Khai phá dữ liệu Web

1.1.1. Tổng quan về khai phá dữ liệu Web

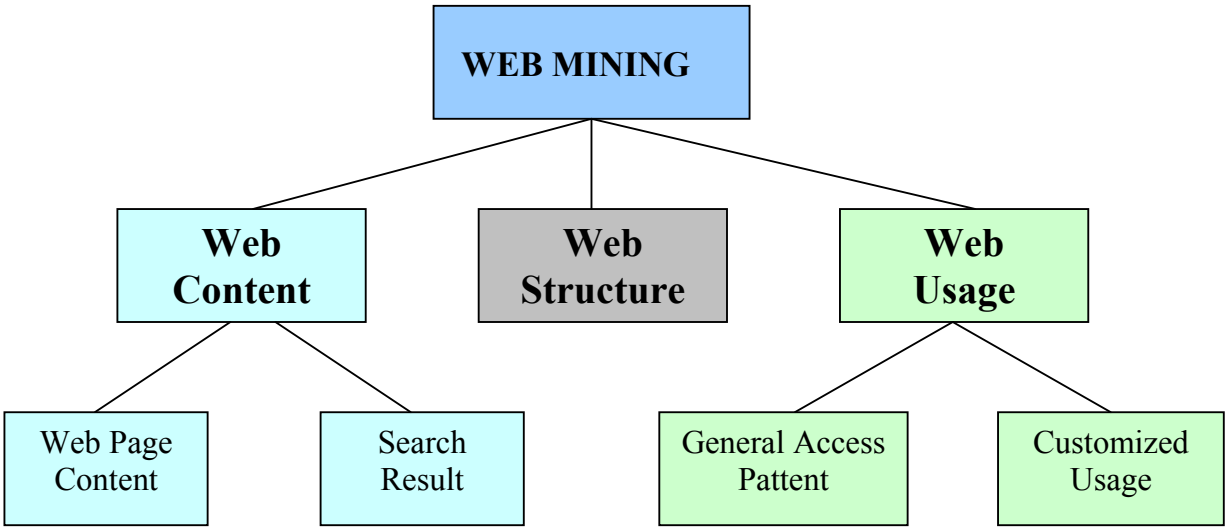
Ngày nay, sự phát triển nhanh chóng của mạng Internet và Intranet đã sinh ra một khối lượng khổng lồ các dữ liệu dạng siêu văn bản (dữ liệu Web). Trong những năm gần đây Intrnet đã trở thành một trong những kênh về khoa học, thông tin kinh tế, thương mại và quảng cáo. Một trong những lý do cho sự phát triển này là chi phí thấp để duy trì một trang Web trên Internet. So sánh với những dịch vụ khác như đăng tin hay quảng cáo trên một tờ báo hay tạp chí, thì một trang Web "đòi" rẻ hơn rất nhiều và cập nhật nhanh chóng hơn tới hàng triệu người dùng khắp mọi nơi trên thế giới. Có thể nói Internet như là cuốn từ điển Bách khoa toàn thư với nội dung và hình thức đa dạng. Nó như một xã hội ảo, nó bao gồm các thông tin về mọi mặt của đời sống kinh tế, xã hội được trình bày dưới dạng văn bản, hình ảnh, âm thanh ...



Hình 1.1: *Khai phá web, công việc không dễ dàng*

Tuy nhiên, Internet là một môi trường đa phương tiện động bao gồm sự kết hợp của các cơ sở dữ liệu không đồng nhất, các chương trình và các giao tiếp người dùng. Rõ ràng, khai phá dữ liệu text chỉ là một lĩnh vực nhỏ trong môi trường này. Khai phá dữ liệu trên Internet, hay thường được gọi là khai phá web ngoài việc cần khai phá được nội dung các trang văn bản, còn phải khai thác được các nguồn lực này cũng như mối quan hệ giữa chúng. Khai phá Web, sự giao thoa giữa khai phá dữ liệu và Word-Wide-Web, đang phát triển mạnh mẽ và bao gồm rất nhiều lĩnh vực nghiên

cứu như trí tuệ nhân tạo, truy xuất thông tin (information retrival) hay các lĩnh vực khác. Các công nghệ Agent-base, truy xuất thông tin dựa trên khái niệm (concept-based), truy xuất thông tin sử dụng case-base reasoning và tính hạng văn bản dựa trên các đặc trưng (features) siêu liên kết... thường được xem là các lĩnh vực nhỏ trong khai phá web. Khai phá Web vẫn chưa được định nghĩa một cách rõ ràng và các chủ đề trong đó vẫn tiếp tục được mở rộng. Tuy vậy, chúng ta có thể hiểu *khai phá web như việc trích ra các thành phần được quan tâm hay được đánh giá là có ích cùng các thông tin tiềm năng từ các tài nguyên hoặc các hoạt động liên quan tới World-Wide Web*[]. Hình 1.2 thể hiện một sự phân loại các lĩnh vực nghiên cứu quen thuộc trong khai phá Web. Người ta thường phân khai phá web thành 3 lĩnh vực chính: khai phá nội dung web (web content mining), khai phá cấu trúc web (web structure mining) và khai phá việc sử dụng web (web usage mining).



Hình 1.2: *Các nội dung trong khai phá Web.*

1.1.2 Các bài toán được đặt ra trong khai phá Web

- Tìm kiếm các thông tin cần thiết: Web quá lớn và quá đa dạng, vì vậy việc tìm được thông tin cần thiết là không đơn giản. Công việc này được giải quyết bởi các máy tìm kiếm.

- Tạo ra các tri thức mới từ các thông tin có sẵn trên Web: Vấn đề này có thể được coi như một vấn đề con của bài toán trên. Ở đây ta mặc định đã có một tập các dữ liệu Web, và ta cần lấy ra được các thông tin hữu ích từ những dữ liệu này.

- Cá nhân hóa các thông tin: Mỗi người dùng thường có các mối quan tâm khác nhau cũng như thích các cách biểu diễn thông tin khác nhau khi tương tác với thế giới Web. Các nghiên cứu về lĩnh vực này sẽ cung cấp các thông tin hữu ích cho những nhà cung cấp thông tin trên Web để họ có thể đạt được mục đích của mình.

- Tìm hiểu về những người tiêu thụ sản phẩm cũng như về cá nhân người dùng: Các nghiên cứu này phục vụ đắc lực để giải quyết vấn đề ở trên. Nó tìm hiểu những điều mà người tiêu dùng muốn và làm. Điều đó sẽ giúp chuyên biệt hóa thông tin cho từng người dùng, giúp thiết kế và quản lý web site một cách hiệu quả, cũng như các vấn đề liên quan tới maketing.

1.1.3 Các lĩnh vực của khai phá dữ liệu Web

1.1.3.1 Khai phá nội dung Web (Web content mining):

Phần lớn các tri thức của World-Wide Web được chứa trong nội dung văn bản. Khai phá nội dung web là các quá trình xử lý để lấy ra các tri thức từ nội dung các trang văn bản hoặc mô tả của chúng. Có hai chiến lược khai phá nội dung web: một là khai phá trực tiếp nội dung của trang web, và một là nâng cao khả năng tìm kiếm nội dung của các công cụ khác như máy tìm kiếm.

- Web Page summarization: liên quan tới việc truy xuất các thông tin từ các văn bản có cấu trúc, văn bản siêu liên kết, hay các văn bản bán cấu trúc. Lĩnh vực này liên quan chủ yếu tới việc khai phá bản thân nội dung các văn bản.

- Search engine result summarization: Tìm kiếm trong kết quả. Trong các máy tìm kiếm, sau khi đã tìm ra những trang Web thoả mãn yêu cầu người dùng, còn một công việc không kém phần quan trọng, đó là phải sắp xếp, chọn lọc kết quả theo mức độ hợp lệ với yêu cầu người dùng. Quá trình này thường sử dụng các thông tin như tiêu đề trang, URL, content-type, các liên kết trong trang web... để tiến hành phân lớp và đưa ra tập con các kết quả tốt nhất cho người dùng.

1.1.3.2. Khai phá cấu trúc web (web structure mining):

Nhờ vào các kết nối giữa các văn bản siêu liên kết, World-Wide Web có thể chứa đựng nhiều thông tin hơn là chỉ các thông tin ở bên trong văn bản. Ví dụ, các liên kết trỏ tới một trang web chỉ ra mức độ quan trọng của trang web đó, trong khi các liên kết đi ra từ một trang web thể hiện các trang có liên quan tới chủ đề đề cập trong trang hiện tại. Và nội dung của khai phá cấu trúc Web là các quá trình xử lý nhằm rút ra các tri thức từ cách tổ chức và liên kết giữa các tham chiếu của các trang web.

1.1.3.3 Khai phá sử dụng web (web usage mining).

Khai phá sử dụng web (web usage mining) hay khai phá hồ sơ web (web log mining) là việc xử lý để lấy ra các thông tin hữu ích trong các hồ sơ truy cập Web. Thông thường các web server thường ghi lại và tích lũy các dữ liệu về các tương tác của người dùng mỗi khi nó nhận được một yêu cầu truy cập. Việc phân tích các hồ sơ truy cập web của các web site khác nhau sẽ dự đoán các tương tác của người dùng khi họ tương tác với Web cũng như tìm hiểu cấu trúc của Web, từ đó cải thiện các thiết kế của các hệ thống liên quan. Có hai xu hướng chính trong khai phá sử dụng web là General Access Pattern Tracking và Customizied Usage tracking.

- General Access Pattern tracking: phân tích các hồ sơ web để biết được các mẫu và các xu hướng truy cập. Các phân tích này có thể giúp cấu trúc lại các site trong các phân nhóm hiệu quả hơn, hay xác định các vị trí quảng cáo hiệu quả nhất, cũng như gắn các quảng cáo sản phẩm nhất định cho những người dùng nhất định để đạt được hiệu quả cao nhất...

- Cusomized Usage tracking: phân tích các xu hướng cá nhân. Mục đích là để chuyên biệt hóa các web site cho các lớp đối tượng người dùng. Các thông tin được hiển thị, độ sâu của cấu trúc site và định dạng của các tài nguyên, tất cả đều có thể chuyên biệt hóa một cách tự động cho mỗi người dùng theo thời gian dựa trên các mẫu truy cập của họ.

1.1.4. Khó khăn

World Wide Web là một hệ thống rất lớn phân bố rộng khắp, cung cấp thông tin trên mọi lĩnh vực khoa học, xã hội, thương mại, văn hóa,... Web là một nguồn tài nguyên giàu có cho Khai phá dữ liệu. Những quan sát sau đây cho thấy Web đã đưa ra những thách thức lớn cho công nghệ Khai phá dữ liệu [1].

1.1.4.1 Web dường như quá lớn để tổ chức thành kho dữ liệu phục vụ Dataming

Các CSDL truyền thống thì có kích thước không lớn lắm và thường được lưu trữ ở một nơi, trong khi đó kích thước Web rất lớn, tới hàng terabytes và thay đổi liên tục, không những thế còn phân tán trên rất nhiều máy tính khắp nơi trên thế giới. Một vài nghiên cứu về kích thước của Web đã đưa ra các số liệu như sau: Hiện nay trên Internet có khoảng hơn một tỷ các trang Web được cung cấp cho người sử dụng., giả sử kích thước trung bình của mỗi trang là 5-10Kb thì tổng kích thước của nó ít nhất là

khoảng 10 terabyte. Còn tỷ lệ tăng của các trang Web thì thật sự gây ấn tượng. Hai năm gần đây số các trang Web tăng gấp đôi và còng tiếp tục tăng trong hai năm tới. Nhiều tổ chức và xã hội đặt hầu hết những thông tin công cộng của họ lên Web. Như vậy việc xây dựng một kho dữ liệu (datawarehouse) để lưu trữ, sao chép hay tích hợp các dữ liệu trên Web là gần như không thể.

1.1.4.2. Độ phức tạp của trang Web lớn hơn rất nhiều so với những tài liệu văn bản truyền thống khác

Các dữ liệu trong các CSDL truyền thống thì thường là loại dữ liệu đồng nhất (về ngôn ngữ, định dạng,…), còn dữ liệu Web thì hoàn toàn không đồng nhất. Ví dụ về ngôn ngữ dữ liệu Web bao gồm rất nhiều loại ngôn ngữ khác nhau (Cả ngôn ngữ diễn tả nội dung lẫn ngôn ngữ lập trình), nhiều loại định dạng khác nhau (Text, HTML, PDF, hình ảnh âm thanh,…), nhiều loại từ vựng khác nhau (Địa chỉ Email, các liên kết (links), các mã nén (zipcode), số điện thoại).

Nói cách khác, trang Web thiếu một cấu trúc thống nhất. Chúng được coi như một thư viện kỹ thuật số rộng lớn, tuy nhiên con số khổng lồ các tài liệu trong thư viện thì không được sắp xếp tuân theo một tiêu chuẩn đặc biệt nào, không theo phạm trù, tiêu đề, tác giả, số trang hay nội dung,... Điều này là một thử thách rất lớn cho việc tìm kiếm thông tin cần thiết trong một thư viện như thế.

1.1.4.3. Web là một nguồn tài nguyên thông tin có độ thay đổi cao

Web không chỉ có thay đổi về độ lớn mà thông tin trong chính các trang Web cũng được cập nhật liên tục. Theo kết quả nghiên cứu [], hơn 500.000 trang Web trong hơn 4 tháng thì 23% các trang thay đổi hàng ngày, và khoảng hơn 10 ngày thì 50% các trang trong tên miền đó biến mất, nghĩa là địa chỉ URL của nó không còn tồn tại nữa. Tin tức, thị trường chứng khoán, các công ty quản cáo và trung tâm phục vụ Web thường xuyên cập nhật trang Web của họ. Thêm vào đó sự kết nối thông tin và sự truy cập bản ghi cũng được cập nhật.

1.1.4.4. Web phục vụ một cộng đồng người dùng rộng lớn và đa dạng

Internet hiện nay nối với khoảng 50 triệu trạm làm việc [1], và cộng đồng người dùng vẫn đang nhanh chóng lan rộng. Mỗi người dùng có một kiến thức, mối quan tâm, sở thích khác nhau. Nhưng hầu hết người dùng không có kiến thức tốt về cấu trúc mạng thông tin, hoặc không có ý thức cho những tìm kiếm, rất dễ bị "lạc" khi

đang "mò mẫm" trong "bóng tối" của mạng hoặc sẽ chán khi tìm kiếm mà chỉ nhận những mảng thông tin không mấy hữu ích.

1.1.4.5. Chỉ một phần rất nhỏ của thông tin trên Web là thực sự hữu ích.

Theo thống kê, 99% của thông tin Web là vô ích với 99% người dùng Web. Trong khi những phần Web không được quan tâm lại bị búi vào kết quả nhận được trong khi tìm kiếm. Vậy thì ta cần phải khai phá Web như thế nào để nhận được trang web chất lượng cao nhất theo tiêu chuẩn của người dùng?

Như vậy chúng ta có thể thấy các điểm khác nhau giữa việc tìm kiếm trong một CSDL truyền thống với vviệc tìm kiếm trên Internet. Những thách thức trên đã đẩy mạnh việc nghiên cứu khai phá và sử dụng tài nguyên trên Internet

1.1.5. Thuận lợi

Bên cạnh những thử thách trên, công việc khai phá Web cũng có những thuận lợi:

1. Web bao gồm không chỉ có các trang mà còn có cả các hyperlink trỏ từ trang này tới trang khác. Khi một tác giả tạo một hyperlink từ trang của ông ta tới một trang A có nghĩa là A là trang có hữu ích với vấn đề đang bàn luận. Nếu trang A càng nhiều Hyperlink từ trang khác trỏ đến chứng tỏ trang A quan trọng. Vì vậy số lượng lớn các thông tin liên kết trang sẽ cung cấp một lượng thông tin giàu có về mối liên quan, chất lượng, và cấu trúc của nội dung trang Web, và vì thế là một nguồn tài nguyên lớn cho khai phá Web.

2. Một máy chủ Web thường đăng ký một bản ghi đầu vào (Weblog entry) cho mọi lần truy cập trang Web. Nó bao gồm địa chỉ URL, địa chỉ IP, timestamp. Dữ liệu Weblog cung cấp lượng thông tin giàu có về những trang Web động. Với những thông tin về địa chỉ URL, địa chỉ IP,… một cách hiển thị đa chiều có thể được cấu trúc nên dựa trên CSDL Weblog. Thực hiện phân tích OLAP đa chiều có thể đưa ra N người dùng cao nhất, N trang Web truy cập nhiều nhất, và khoảng thời gian nhiều người truy cập nhất, xu hướng truy cập Web.

1.2 Tổng quan về máy tìm kiếm

1.2.1 Nhu cầu

Như đã đề cập ở phần trên, Internet là một kho thông tin khổng lồ và phức tạp. Thông tin trên các trang Web đa dạng về mặt nội dung cũng như hình thức. Tuy nhiên

cùng với sự đa dạng và số lượng lớn thông tin như vậy đã nảy sinh vấn đề quá tải thông tin. Cùng với sự thay đổi và phát triển hàng ngày hàng giờ về nội dung cũng như số lượng của các trang Web trên Internet thì vấn đề tìm kiếm thông tin đối với người sử dụng lại ngày càng khó khăn. Đối với mỗi người dùng chỉ một phần rất nhỏ thông tin là có ích, chẳng hạn có người chỉ quan tâm đến trang Thể thao, Văn hóa mà không mấy khi quan tâm đến Kinh tế. Người ta không thể tìm tự kiếm địa chỉ trang Web chứa thông tin mà mình cần, do vậy đòi hỏi cần phải có một trình tiện ích quản lý nội dung của các trang Web và cho phép tìm thấy các địa chỉ trang Web có nội dung giống với yêu cầu của người tìm kiếm.

Định nghĩa []:Máy tìm kiếm (search engine) là một hệ thống được xây dựng nhằm tiếp nhận các yêu cầu tìm kiếm của người dùng (thường là một tập các từ khóa), sau đó phân tích yêu cầu này và tìm kiếm thông tin trong cơ sở dữ liệu được tải xuống từ Web và đưa ra kết quả là các trang web có liên quan cho người dùng.

Cụ thể, người dùng gửi một truy vấn, dạng đơn giản nhất là một danh sách các từ khóa, và máy tìm kiếm sẽ làm việc để trả lại một danh sách các trang Web có liên quan hoặc có chứa các từ khóa đó. Phức tạp hơn, thì truy vấn là cả một văn bản hoặc một đoạn văn bản hoặc nội dung tóm tắt của văn bản. Một số máy tìm kiếm điển hình hiện nay: Yahoo, Google, Alvista,...

1.2.2 Cơ chế hoạt động của máy tìm kiếm.

Một máy tìm kiếm có thể được xem như là một ví dụ của hệ thống truy xuất thông tin Information Retrival (IR). Một hệ thống truy xuất thông tin IR thường tập trung vào việc cải thiện hiệu quả thông tin được lấy ra bằng cách sử dụng việc đánh chỉ số dựa trên các từ khóa (term-base indexing)[8,11] và kỹ thuật tổ chức lại các câu truy vấn (query refomulation technique)[32]. Quá trình xử lý các văn bản dựa trên từ khóa ban đầu trích ra các từ khóa trong văn bản sử dụng một từ điển được xây dựng trước, một tập các từ dừng, và các qui tắc (stemming rule)[10] để chuyển các hình thái của từ về dạng từ gốc. Sau khi các từ khóa đã được lấy ra, và thường sử dụng phương pháp TF-IDF (hoặc biến thể của nó) [31,33] để xác định mức độ quan trọng của các từ khóa. Do đó, một văn bản có thể được biểu diễn bởi một tập các từ khóa và độ quan trọng của chúng. Mức độ tương tự đo được giữa một câu truy vấn và một văn bản chính bằng tích trực tiếp tích direct product giữa hai vector các từ khóa tương ứng. Để thể hiện mức độ hợp lệ của các văn bản và câu truy vấn, các văn bản được lấy ra được