Chapter 1

Usage Mining for and on the Semantic Web

Gerd Stumme[*], *Andreas Hotho*[*], *Bettina Berendt*[×]
[*]Institute for Applied Computer Science and Formal Description Methods (AIFB), University of Karlsruhe, D-76128 Karlsruhe, Germany, www.aifb.uni-karlsruhe.de/WBS/gst
[×]Institute of Information Systems, Humboldt University Berlin, Spandauer Str. 1, D-10178 Berlin, Germany, berendt@wiwi.hu-berlin.de

Abstract:
Semantic Web Mining aims at combining the two fast-developing research areas Semantic Web and Web Mining. Web Mining aims at discovering insights about the meaning of Web resources and their usage. Given the primarily syntactical nature of data Web mining operates on, the discovery of meaning is impossible based on these data only. Therefore, formalizations of the semantics of Web resources and navigation behavior are increasingly being used. This fits exactly with the aims of the Semantic Web: the Semantic Web enriches the WWW by machine-processable information which supports the user in his tasks. In this paper, we discuss the interplay of the Semantic Web with Web Mining, with a specific focus on usage mining.

Keywords: Web mining, Semantic Web, ontologies

1.1 Introduction

Web Usage Mining is the application of data mining methods to the analysis of recordings of Web usage, most often in the form of Web server logs. One of its central problems is the large number of patterns that are usually found: among these, how can the *interesting* patterns be identified?

For example, an application of association rule analysis to a Web log will typically return many patterns like the observation that 90% of the users who made a purchase in an online shop also visited the homepage—a pattern that is trivial because the homepage is the site's main entry point.

Statistical measures of pattern quality like support and confidence, and measures of interestingness based on the divergence from prior beliefs are a primarily syntactical approach to this problem. They need to be complemented by an understanding of what a site and its usage patterns are about, i.e. a semantic approach.

A popular approach for modelling sites and their usage is related to OLAP techniques: a modelling of the pages in terms of (possibly multiple) concept hierarchies, and an investigation of patterns at different levels of abstraction, i.e. a knowledge discovery cycle which iterates over various "roll-ups" and "drill-downs". Concept hierarchies conceptualize a domain in terms of taxonomies such as product catalogs, topical thesauri, etc. The expressive power of this form of knowledge representation is limited to is-a relationships. However, for many applications, a more expressive form of knowledge representation is desirable, for example *ontologies* that allow arbitrary relations between concepts.

A second problem facing many current analyses that take semantics into account is that the conceptualizations often have to be hand-crafted to represent a site that has grown independently of an overall conceptual design, and that the mapping of individual pages to this conceptualization may have to be established.

It would thus be desirable to have a rich semantic model of a site, of its content and its (hyperlink) structure, a model that captures the complexity of the manifold relationships between the concepts covered in a site, and a model that is "built into" the site in the sense that the pages requested by visitors are directly associated with the concepts and relations treated by it.

The *Semantic Web* is just this: today's Web enriched by a formal semantics in form of ontologies that captures the meaning of pages and links in a machine-understandable form. The main idea of the Semantic Web is to enrich the current Web by machine-processable information in order to allow for semantic-based tools supporting the human user. In this paper, we discuss on one hand how the Semantic Web can improve Web usage mining, and on the other hand how usage mining can be used to built up the Semantic Web.

After a short overview of the relevant areas of Web Mining in the next section, Section 1.3 will describe this understanding of the Semantic Web in more detail. Section 1.4 then gives an overview of how semantics can enhance Web usage mining, ranging to a mining of the Semantic Web itself. We also discuss how to incorporate knowledge about behavior in Web hypermedia that transcends the semantics of sites in themselves.

Section 1.5 then illustrates how mining can contribute to the development of the Semantic Web by automatically extracting knowledge from Web resources.

We use the term *Semantic Web Mining* to denote these various methods of mining the Semantic Web and mining for the Semantic Web. Beside Semantic Web Usage Mining, Semantic Web Mining also includes Semantic Web Content and Structure Mining. We discussed the overall view of Semantic Web Mining in [5]. In this paper, we will focus on more specific aspects of Semantic Web Usage Mining.

Ideally, all methods addressed above should be combined to go from a site and its usage to its semantics and back. In the conclusion, we sketch a phase model of this feedback loop, which will improve the understanding of the site and its usage, for the purpose of improving the site for its users.

1.2 Web (Usage) Mining

Web mining is the application of data mining techniques to the content, structure, and usage of Web resources. This can help to discover global as well as local structure within and between Web pages. Like other data mining applications, Web mining can profit from given structure on data (as in database tables), but it can also be applied to semi-structured or unstructured data like free-form text. This means that Web mining is an invaluable help in the transformation from human understandable content to machine understandable semantics.

A distinction is generally made between Web mining that operates on the Web resources themselves (often further differentiated into *content* and *structure* mining), and mining that operates on visitors' *usage* of these resources.

Web content mining is a form of text mining (for recent overviews, see [9, 42]). It concentrates on the content of individual pages, which is contained in the HTML, script, etc. code that generates a page. It can therefore take advantage of the semi-structured nature of these types of text. The hyperlink structure between those pages is, on the one hand, a structure over and above the individual page. This is utlized in *Web structure mining* approaches like the Google PageRank algorithm that determines the relevance of a page by how many other pages "cite" it [40], see also [29]. On the other hand, hyperlinks are part of the textual content of a page. This is particularly true for pages in which hyperlinks, like other elements, are (semantically) marked up. In the following, we will therefore treat these two areas in an integrated fashion [14]. Web content/structure mining can be used, among other things, to extract information from Web pages to determine keywords decribing content, or even to assign Web pages to a domain model, to detect events or track developments in time-varying series of Web resources like newswire articles. These techniques, and their application for understanding Web usage, will be discussed in more detail in section 1.5.

In *Web usage mining*, the primary Web resource that is being mined is a record of the requests made by visitors to a Web site, most often collected in a Web server log [47]. The content and structure of Web pages, and in particular those of one Web site, reflect the intentions of the authors and designers of the pages and the underlying information architecture. The actual behavior of the users of these resources may reveal additional structure.

First, relationships may be induced by usage where no particular structure was designed. For example, in an online catalog of products, there is usually either no inherent

structure (different products are simply viewed as a set), or one or several hierarchical structures given by product categories, manufacturers, etc. Mining the visits to that site, however, one may find that many of the users who were interested in product A were also interested in product B. Here, "interest" may be measured by requests for product description pages, or by the placement of that product into the shopping cart (indicated by the request for the respective pages). The identified association rules are at the center of cross-selling and up-selling strategies in E-commerce sites: When a new user shows interest in product A, she will receive a recommendation for product B (cf. [35, 31]).

Second, relationships may be induced by usage where a different relationship was intended. For example, sequence mining may show that many of the users who visited page C later went to page D, along paths that indicate a prolonged search (frequent visits to help and index pages, frequent backtracking, etc.) [13, 25]. This can be interpreted to mean that visitors wish to reach D from C, but that this was not foreseen in the information architecture, hence that there is at present no hyperlink from C to D. This insight can be used for static site improvement for all users (adding a link from C to D), or for dynamic recommendations personalized for the subset of users who go to C ("you may wish to also look at D").

It is useful to combine Web usage mining with content and structure analysis in order to "make sense" of observed frequent paths and the pages on these paths. This can be done using a variety of methods. Many of these methods rely on a mapping of pages into an ontology. And underlying ontology and the mapping of pages into it may already be available, the mapping of pages into an existing ontology may need to be learned, and/or the ontology itself may have to be inferred first.

In the following sections, we will first investigate the notions of semantics (as used in the Semantic Web) and ontologies in more detail. We will then look at how the use of ontologies, and other ways of identifying the meaning of pages, can help to make Web Mining go semantic. Lastly, we will investigate how ontologies and their instances can be learned.

1.3 Semantic Web

The Semantic Web is based on a vision of Tim Berners-Lee, the inventor of the WWW. The great success of the current WWW leads to a new challenge: a huge amount of data is interpretable by humans only; machine support is limited. Berners-Lee suggests to enrich the Web by machine-processable information which supports the user in his tasks. For instance, today's search engines are already quite powerful, but still return too often too large or inadequate lists of hits. Machine-processable information can point the search engine to the relevant pages and can thus improve both precision and recall.

For instance, it is today almost impossible to retrieve information with a keyword search when the information is spread over several pages. Consider, e. g., the query for web mining experts in a company intranet, where the only explicit information stored are the relationships between people and the courses they attended on one hand, and between courses and the topics they cover on the other hand. In that case, the use of

a rule stating that people who attended a course which was about a certain topic have knowledge about that topic might improve the results.

The process of building the Semantic Web is today still heavily going on. Its structure has to be defined, and this structure has then to be filled with life. In order to make this task feasible, one should start with the simpler tasks first. The following steps show the direction where the Semantic Web is heading:

1. Providing a common syntax for machine understandable statements.
2. Establishing common vocabularies.
3. Agreeing on a logical language.
4. Using the language for exchanging proofs.

Berners-Lee suggested a layer structure for the Semantic Web: (i) Unicode/URI, (ii) XML/Name Spaces/ XML Schema, (iii) RDF/RDF Schema, (iv) Ontology vocabulary, (v) Logic, (vi) Proof, (vii) Trust.[1] This structure reflects the steps listed above. It follows the understanding that each step alone will already provide added value, so that the Semantic Web can be realized in an incremental fashion.

On the first two layers, a common syntax is provided. *Uniform resource identifiers (URIs)* provide a standard way to refer to entities,[2] while *Unicode* is a standard for exchanging symbols. The *Extensible Markup Language (XML)* fixes a notation for describing labeled trees, and XML Schema allows to define grammars for valid XML documents. XML documents can refer to different *namespaces* to make explicit the context (and therefore meaning) of different tags. The formalizations on these two layers are nowadays widely accepted, and the number of XML documents is increasing rapidly.

The *Resource Description Framework (RDF)* can be seen as the first layer which is part of the Semantic Web. According to the W3C recommendation [37], RDF "is a foundation for processing metadata; it provides interoperability between applications that exchange machine-understandable information on the Web." RDF documents consist of three types of entities: resources, properties, and statements. Resources may be web pages, parts or collections of web pages, or any (real-world) objects which are not directly part of the WWW.

In RDF, resources are always addressed by URIs. Properties are specific attributes, characteristics, or relations describing resources. A resource together with a property having a value for that resource form an RDF statement. A value is either a literal, a resource, or another statement. Statements can thus be considered as object–attribute–value triples.

The middle part of Figure 1.1 shows an example of RDF statements. Two of the authors of the present paper (i. e., their Web pages) are represented as resources 'URI-GST' and 'URI-AHO'. The statement on the lower right consists of the resource 'URI-AHO' and the property 'cooperateswith' with the value 'URI-GST' (which again is a

[1] see http://www.w3.org/DesignIssues/Semantic.html

[2] *URL (uniform resource locator)* refers to a locatable URI, e.g. an http://... address. It is often used as a synonym, although strictly speaking URLs are a subclass of URIs, see http://www.w3.org/Addressing.

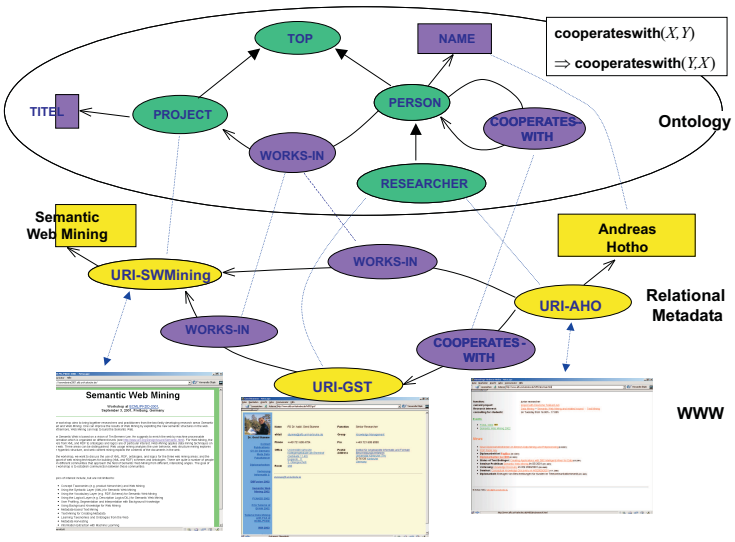

Figure 1.1: The relation between the WWW, relational metadata, and ontologies.

resource). The resource 'URI-SWMining' has as value for the property 'title' the literal 'Semantic Web Mining'.

The data model underlying RDF is basically a directed labeled graph. RDF Schema defines a simple modeling language on top of RDF which includes classes, is-a relationships between classes and between properties, and domain/range restrictions for properties. RDF and RDF Schema are noted with the syntax of XML, but they do not employ the tree semantics of XML.

The next layers are the *ontology vocabulary* and *logic*. Today the Semantic Web community considers these levels rather as one single level as most ontologies allow for logical axioms. Following [18], an ontology is "an explicit formalization of a shared understanding of a conceptualization". This high-level definition is realized differently by different research communities. However, most of them have a certain understanding in common, as most of them include a set of concepts, a hierarchy on them, and relations between concepts. Most of them also include axioms in some specific logic. To give a flavor, we present here just the core of our own definition [50, 7], as it is reflected by the Karlsruhe Ontology framework KAON.[3] It is built in a modular way, so that different needs can be fulfilled by combining parts.

Definition 1.3.1 *A* core ontology with axioms *is a tuple* $\mathcal{O} := (C, \leq_C, R, \sigma, \leq_R, \mathcal{A})$ *consisting of*

- *two disjoint sets* C *and* R *whose elements are called* concept identifiers *and* relation identifiers*, resp.,*

[3] http://kaon.semanticweb.org

- *a partial order $\leq_C$ on C, called* concept hierarchy *or* taxonomy*,*
- *a function $\sigma: R \rightarrow C^+$ called* signature *(where C^+ is the set of all finite tuples of elements in C),*
- *a partial order $\leq_R$ on R, called* relation hierarchy*, where $r_1 \leq_R r_2$ implies $|\sigma(r_1)| = |\sigma(r_2)|$ and $\pi_i(\sigma(r_1)) \leq_C \pi_i(\sigma(r_2))$, for each $1 \leq i \leq |\sigma(r_1)|$, with π_i being the projection on the ith component, and*
- *a set $\mathcal{A}$ of logical axioms in some logical language $\mathcal{L}$.*

This definition constitutes a core structure that is quite straightforward, well-agreed upon, and that may easily be mapped onto most existing ontology representation languages. Step by step the definition can be extended by taking into account axioms, lexicons, and knowledge bases [50].

As an example, have a look at the top of Figure 1.1. The set C of concepts is the set {Top, Project, Person, Researcher, Literal}, and the concept hierarchy $\leq_C$ is indicated by the arrows with a bold head. The set R of relations is the set {works-in, researcher, cooperates-with, name}. The relation 'worksin' has (Person, Project) as signature, the relation 'name' has (Person, Literal) as signature.[4] In this example, the hierarchy on the relations is flat, i. e., $\leq_R$ is just the identity relation. For an example of a non-flat relation, have a look at Figure 1.2. The objects of the metadata level can be seen as instances of the ontology concepts. For example, 'URI-SWMining' is an instance of the concept 'Project', and thus by inheritance also of the concept 'Top'. Up to here, RDF Schema would be sufficient for formalizing the ontology.

Often ontologies contain also logical axioms. By applying logical deduction, one can then infer new knowledge from the information which is stated implicitly. The axiom in Figure 1.1 states for instance that the 'cooperates-with' relation is symmetric. From it, one can logically infer that the person addressed by 'URI-AHO' is cooperating with the person addressed by 'URI-GST' (and not only the other way around).

A priori, any knowledge representation mechanism[5] can play the role of a Semantic Web language. *Frame Logic* (or *F–Logic*; [27]), for instance, provides a semantically founded knowledge representation based on the frame and slot metaphor. Probably the most popular framework at the moment are Description Logics (DL). DLs are subsets of first order logic which aim at being as expressive as possible while still being decidable. The description logic $\mathcal{SHIQ}$ provides the basis for the web language DAML+OIL.[6] Its latest version is currently established by the W3C Web Ontology Working Group (WebOnt)[7] under the name OWL.

Several tools are in use for the creation and maintenance of ontologies and metadata, as well as for reasoning within them. Our group has developed *OntoEdit* [51, 52], an ontology editor which is connected to *Ontobroker* [17], an inference engine for F–Logic. It provides means for semantic based query handling over distributed resources.

[4] It is a drawing convention to have relations with Literal as range drawn this way, as they are sometimes considered as *attributes*.

[5] See [48] for a general discussion.

[6] http://www.daml.org

[7] http://www.w3.org/2001/sw/WebOnt/

Recently, the Karlsruhe Ontology Framework KAON [7] has been set up as follow–up of OntoEdit. It is an open-source ontology management infrastructure targeted for business applications consisting of a comprehensive tool suite. KAON allows for easy ontology creation and management, as well as building ontology-based applications. A connection to the FaCT[8] inference engine for the Description Language $\mathcal{SHIQ}$ is planned.

Proof and *trust* are the remaining layers. They follow the understanding that it is important to be able to check the validity of statements made in the (Semantic) Web. These two layers are rarely tackled today, but are interesting topics for future research. In this paper, we will focus our interest on the XML, RDF, ontology and logic layers.

1.4 Using Semantics for Usage Mining and Mining the Usage of the Semantic Web

Semantics can be utilized for Web Mining for different purposes. Some of the approaches presented in this section rely on a comparatively *ad hoc* formalization of semantics, while others exploit the full power of the Semantic Web. The Semantic Web offers a good basis to enrich Web Mining: The types of (hyper)links are now described explicitly, allowing the knowledge engineer to gain deeper insights in Web structure mining; and the contents of the pages come along with a formal semantics, allowing her to apply mining techniques which require more structured input. Because the distinction between the use of semantics for Web mining and the mining of the Semantic Web itself is all but sharp, we will discuss both in an integrated fashion.

Web usage mining benefits from including semantics into the mining process for the simple reason that the application expert as the end user of mining results is interested in *events in the application domain*, in particular user behavior, while the data available—Web server logs—are technically oriented sequences of *HTTP requests*.A central aim is therefore to map HTTP requests to meaningful units of application events.

Application events are defined with respect to the application domain and the site, a non-trivial task that amounts to a detailed formalization of the site's business model. For example, relevant E-business events include product views and product click-throughs in which a user shows specific interest in a specific product by requesting more detailed information (e.g., from the Beach Hotel to a listing of its prices in the various seasons).[9] Related events are click-throughs to a product category (e.g., from the Beach Hotel to the category of all Wellness Hotels); or click-through from a banner ad. Further product-oriented events include shopping cart changes and product purchases or bids.

Web server logs generally contain at least some information on an event that was marked by the user's request for a specific Web page, or the system's generating a page

[8] http://www.cs.man.ac.uk/~horrocks/FaCT

[9] This tourism example, used throughout the paper, is inspired by the Getess project (http://www.getess.de/index_en.html), which provides ontology-based access to tourism Web pages in Mecklenburg-Vorpommern (http://www.all-in-all.de/), a region in north-eastern Germany.

to acknowledge the successful completion of a transaction. For example, consider a tourism Web site that allows visitors to search hotels according to different criteria, to look at detailed descriptions of these hotels, to make reservations, and so on. In the Web site, a hotel room reservation event may be identified by the recorded delivery of the page `reserve.php?user=12345&hotel=Beach_Hotel&people=2&arrive=01May&depart=04May`, which was generated after the user chose "room for 2 persons" in the "Beach Hotel" and typed in the arrival and departure dates of his desired stay. What information the log contains, and whether this is sufficient, will depend on the technical set-up of the site as well as on the purposes of the analysis.

So what are the aspects of application events that need to be reconstructed using semantics? In the following sections, we will show that a requested Web page is, first, about some *content*, second, the request for a specific *service* concerning that content, and third, usually part of a larger sequence of events. We will refer to the first two as *atomic application events*, and to the third as *complex application event*.

1.4.1 Atomic application events: Content

A requested Web page is about something, usually a product or other object described in the page. For example, `search_hotel.html?facilities=tennis` may be a page about hotels, more specifically a listing of hotels, with special attention given to a detailed indication of their sports facilities. To describe content in this way, URLs are generally mapped to concepts. The concepts are usually organized in taxonomies (also called "concept hierarchies", see [19] and the definition in Section 1.3). For example, a `tennis` court is a `facility`. Introducing relations, we note that a `facility` `belongs-to` an `accommodation`, etc. (see Fig. 1.2).

1.4.2 Atomic application events: Service

A requested Web page reflects a purposeful user activity, often the request for a specific service. For example, `search_hotel.html?facilities=tennis` was generated after the user had initiated a search by hotel facilities (stating `tennis` as the desired value). This way of analyzing requests gives a better sense of what users wanted and expected from the site, as opposed to what they received in terms of the eventual content of the page.

To a certain extent, the requested service is associated with the request's URL stem and the delivered page's content (e.g., the URL `search_hotel.html` says that the page was a result of a search request). However, the delivered page's content may also be meaningless for the understanding of user intentions, as is the case when the delivered page was a "404 File not found".

More information is usually contained in the specifics of the user query that led to the creation of the page. This information may be contained in the URL query string, which is recorded in the Web server log if the common request method GET is used. The query string may also be recorded by the application server in a separate log.

As an example, we have used an ontology to describe a Web site which operates on relational databases and also contains a number of static pages, together with an automated classification scheme that relies on mapping the query strings for dynamic page

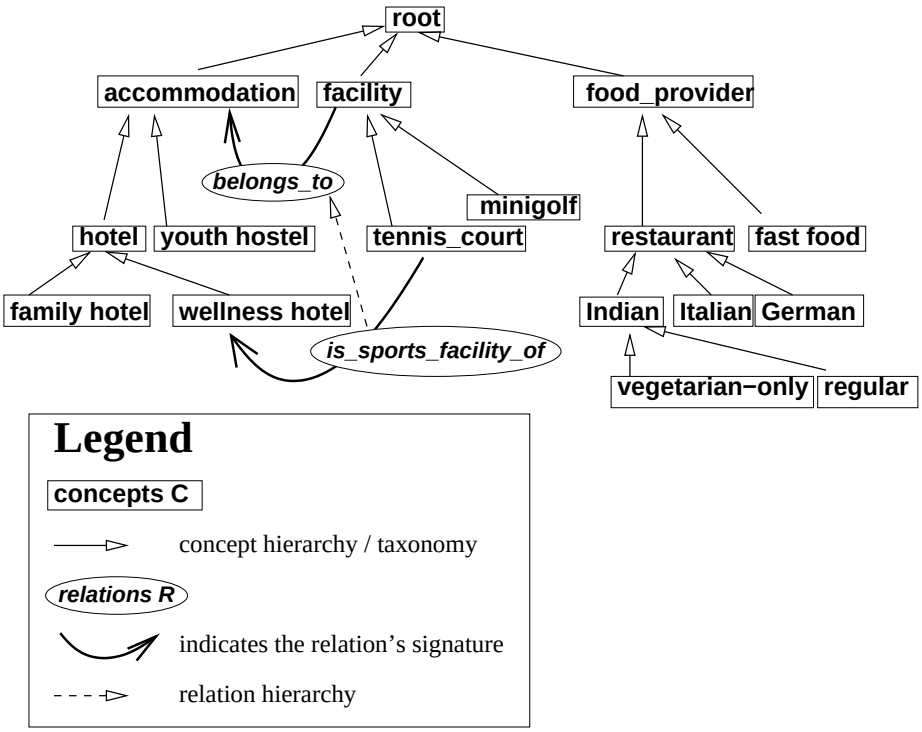

Figure 1.2: Parts of the ontology of the content of a fictitious tourism Web site.

generation to concepts [4]. Pages are classified according to multiple concept hierarchies that reflect content (type of object that the page describes), structure (function of pages in object search), and service (type of search functionality chosen by the user). A path can then be regarded as a sequence of (more or less abstract) concepts in a concept hierarchy, allowing the analyst to identify strategies of search. This classification can make Web usage mining results more comprehensible and actionable for Web site redesign or personalization: The semantic analysis has helped to improve the design of search options in the site, and to identify behavioral patterns that indicate whether a user is likely to successfully complete a search process, or whether he is likely to abandon the site [44]. The latter insights could be used to dynamically generate help messages for new users.

Oberle [39] develops a scheme for application server logging of user queries with respect to a full-blown ontology (a "knowledge portal" in the sense of [21]). This allows the analyst to utilize the full expressiveness of the ontology language, which enables a wide range of inferences going beyond the use of taxonomy-based generalizations. He gives examples of possible inferences on queries to a community portal, which can help support researchers in finding potential cooperation partners and projects. A large-scale evaluation of the proposal is under development.

The ontologies of content and services of a Web site as well as the mapping of

pages into them may be obtained in various ways. At one extreme, ontologies may be hand-crafted ex post; at the other extreme, they may be the generating structure of the Web site (in which case also the mapping of pages to ontology elements is already available). In most cases, mining methods themselves must be called upon to establish the ontology (*ontology learning*) and/or the mapping (*instance learning*), for example by using methods of learning relations (e.g., [32]) and information extraction (e.g., [15, 30]).

1.4.3 Complex application events

A requested Web page, or rather, the activity/ies behind it, is generally part of a more extended behavior. This may be a problem-solving strategy consciously pursued by the user (e.g., to narrow down search by iteratively refining search terms), a canonical activity sequence pertaining to the site type (e.g., catalog search/browse, choose, add-to-cart, pay in an E-commerce setting [34]), or a description of behavior identified by application experts in exploratory data analysis. An example of the latter is the distinction of four kinds of online shopping strategies by [38]: *directed buying*, *search/deliberation*, *hedonic browsing*, and *knowledge building*. The first group is characterized by focused search patterns and immediate purchase. The second is more motivated by a future purchase and therefore tends to browse through a particular category of products rather than directly proceed to the purchase of a specific product. The third is entertainment- and stimulus-driven, which occasionally results in spontaneous purchases. The fourth also shows exploratory behavior, but for the primary goal of information acquisition as a basis for future purchasing decisions. Moe characterized these browsing patterns in terms of product and category pages visited on a Web site. Spiliopoulou, Pohle, and Teltzrow [45] transferred this conceptualization to the analysis of a non-commercial information site. They formulated regular expressions that capture the behavior of search/deliberation and knowledge building, and used sequence mining to identify these behaviors in the site's logs.

The ontologies required in this case are most often sequences or, more generally, regular expressions whose atoms are activity and/or content units as described in "Service" and "Content" above. In general, the sequential order within a behavior will be relevant; thus sequence mining is applied. Figure 1.3 illustrates these concepts using the running example.

1.4.4 How is knowledge about application events used in mining?

Once requests have been mapped to concepts, the question arises how knowledge is gained from these transformed data. We will investigate the treatment of atomic and of complex application events in turn.

Mining using multiple taxonomies is related to OLAP data cube techniques: objects (in this case, requests or requested URLs) are described along a number of dimensions, and concept hierarchies or lattices are formulated along each dimension to allow more abstract views (cf. [54, 28, 22, 49]).

The analysis of data abstracted using taxonomies is crucial for many mining applications to generate meaningful results: In a site with dynamically generated pages,

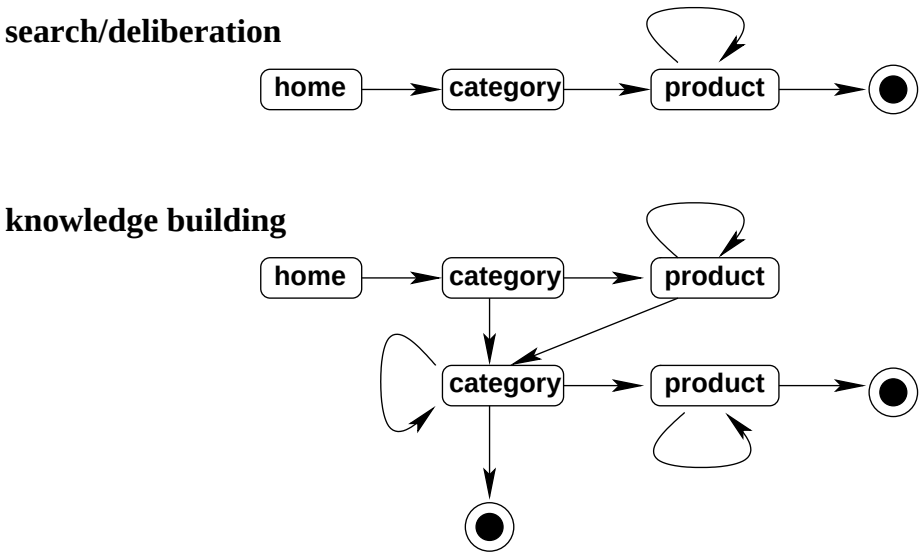

Figure 1.3: Parts of the ontology of the complex application events of the example site.

each individual page will be requested so infrequently that no regularities may be found in an analysis of navigation behavior. Rather, regularities may exist at a more abstract level, leading to rules like "visitors who stay in Wellness Hotels also tend to eat in restaurants". Second, patterns mined in past data are not helpful for applications like recommender systems when new items are introduced into product catalog and/or site structure: The new Pier Hotel cannot be recommended simply because it was not in the tourism site before and thus could not co-occur with any other item, be recommended by another user, etc. A knowledge of regularities at a more abstract level could help to derive a recommendation of the Pier Hotel because it too is a Wellness Hotel (and there are criteria for recommending Wellness Hotels).

After the preprocessing steps in which access data have been mapped into taxonomies, two main approaches are taken in subsequent mining steps. In many cases, mining operates on concepts at a chosen level of abstraction. For example, the sessions are transformed into points in a feature space [36], or on the sessions transformed into sequences of content units at a given level of description (for example, association rules can be sought between abstract concepts such as Wellness Hotels, tennis courts, and restaurants). This approach is usually combined with interactive control of the software, so that the analyst can re-adjust the chosen level of abstraction after viewing the results (e.g., in the miner WUM; see [4] for a case study).

Alternatively to this 'static' approach, other algorithms identify the most specific level of relationships by choosing concepts dynamically. This may lead to rules like "People who stay at Wellness Hotels tend to eat at vegetarian-only Indian restaurants"—linking hotel-choice behavior at a comparatively high level of abstraction with restaurant-choice behavior at a comparatively detailed level of description.

For example, Srikant and Agrawal [46] search for associations in given taxonomies,

using support and confidence thresholds to guide the choice of level of abstraction. Dai and Mobasher [16] present a scheme for aggregating towards more general concepts when an explicit taxonomy is missing. They apply clustering to sets of sessions; this clustering identifies related concepts at different levels of abstraction.

Semantic Web Usage Mining for complex application events involves two steps of mapping requests to events. As discussed in Section 1.4.3 above, complex application events are usually defined by regular expressions in atomic application events (at some given level of abstraction in their respective hierarchies). Therefore, in a first step, URLs are mapped to atomic application events at the required level of abstraction. In a second step, a sequence miner can then be used to discover sequential patterns in the transformed data. The shapes of sequential patterns sought, and the mining tool used, determine how much prior knowledge can be used to constrain the patterns identified: They range from largely unconstrained first-order or k-th order Markov chains [6], to regular expressions that specify the atomic activities completely (the name of the concept) or partially (a variable matching a set of concepts) [43, 2].

Examples of the use of regular expressions describing application-relevant courses of events include search strategies [4], a segmentation of visitors into customers and non-customers [44], and a segmentation of visitors into different interest groups based on the customer buying cycle model from marketing [45].

To date, few commonly agreed-upon models of Semantic Web behavior exist. The still largely exploratory nature of the field implies that highly interactive data preparation and mining tools are of paramount importance: They give the best support for domain experts working with analysts to contribute their background knowledge in an iterative mining cycle. A central element of interactive tools for exploration is visualization. Visualization allows the detection of further structure in sequential patterns (for a survey, see [3]).

However, with the increasing standardization of many Web applications, and the increasing confluence of mining research with application domain research (e.g., marketing), the number of standard courses of events is likely to grow. Examples are the predictive schemes of E-commerce sites (see the example from [34] mentioned in Section 1.4.3 above), and the description of browsing strategies given by [38].

1.5 Extracting Semantics from Web Usage

The effort behind the Semantic Web is to add semantic annotation to Web documents in order to access knowledge instead of unstructured material. The purpose is to allow knowledge to be managed in an automatic way. Web Mining can help to learn definitions of structures for knowledge organization (e. g., ontologies) and to provide the population of such knowledge structures.

All approaches discussed here are semi-automatic. They assist the knowledge engineer in extracting the semantics, but cannot completely replace her. In order to obtain high-quality results, one cannot replace the human in the loop, as there is always a lot of tacit knowledge involved in the modeling process [8]. A computer will never be able to fully consider background knowledge, experience, or social conventions. If this were the case, the Semantic Web would be superfluous, since then machines like search

engines or agents could operate directly on conventional Web pages. The overall aim of our research is thus not to replace the human, but rather to provide him with more and more support.

In [5], we have discussed how content, structure, and usage mining can be used for creating Semantics. Here we focus on the contribution of usage mining.

In the World Wide Web as in other places, much knowledge is socially constructed. This social behavior is reflected by the usage of the Web. One tenet related to this view is that navigation is not only driven by formalized relationships or the underlying logic of the available Web resources, but that it "is an information browsing strategy that takes advantage of the behavior of like-minded people" ([10, p.18]). Recommender systems based on "collaborative filtering" have been the most popular application of this idea. In recent years, the idea has been extended to consider not only ratings, but also Web usage as a basis for the identification of like-mindedness ("People who liked/bought this book also looked at ..."); see [35] for a recent mining-based system; see also [24] for a classic, although not mining-based, application.

Web usage mining by its definition always creates patterns that structure pages in some way. An example is the "relatedness" of different products that frequently co-occur in user visits and can therefore be used for cross-selling and up-selling recommendations. The question is how usage can create patterns that help build the Semantic Web.

Put differently, the question is how usage patterns can reveal semantic relationships better than the automated analysis of collections of single pages alone may do. For Semantic Web Usage Mining, this means that some content may actually be defined only after and through (frequent) usage. An illustrative selection of approaches will highlight the research questions in this current area of research.

Ypma and Heskes [53] model navigation in terms of hidden Markov models, with the hidden states being page categories, and the observed request events being instances of them. Their main aim was to show that a meaningful page categorization may be learned simultaneously with the user labeling and inter-category transitions; "semantic labels" (such as "sports" pages) were assigned to a state afterwards by manual inspection of the pages instantiating that state.

Chi et al. [12, 11] identify frequent paths through a site. Based on the keywords extracted from the pages along the path, they compute the likely "information scent" followed, i.e. the intended goal of the path. The information scent is a set of weighted keywords, which can be manually inspected using an interactive, highly visual tool, and then labeled more concisely. Thus, usage creates a set of "information goals" users expect from the site.

User navigation has been employed to infer topical relatedness, i.e. the relatedness of a set of pages to a topic as given by the terms of a query to a search engine. Aggarwal [1] refers to this as "collaborative crawling" and proposes it as a method for improving on "focused" and "intelligent" crawling which uses only the information from page content and hyperlink structure. A classification of pages into "satisfying the user defined predicate" and "not satisfying the predicate" is thus learned from usage, structure, and content information. An obvious application is to mine user navigation to improve search engine ranking [23, 26].

Many approaches use a combination of content extraction techniques (like keyword

analysis) and techniques for extracting patterns of usage to generate recommendations. For example, in content-based collaborative filtering, textual categorization of documents are used for generating pseudo-rankings for every user-document pair [33]. [41] use a combination of IE techniques analyzing single pages, ontologies, and the mining of a user's previous search history to make recommendations for query improvement in a search engine. The basic idea is (a) to offer terms that are shown in the hierarchy as related, and (b) to infer from terms that occurred frequently in previous search histories a relative weighting on the set of pages that are described only coarsely by the few terms of the initial current query.

To sum up, these examples show various ways of how the Semantic Web can be built up using techniques of Web (usage) mining.

1.6 Conclusion and Outlook

In this paper, we have studied the combination of the two fast-developing research areas Semantic Web and Web Mining, especially usage mining. We discussed how Semantic Web Usage Mining can improve the results of 'classical' usage mining by exploiting the new semantic structures in the Web; and how the construction of the Semantic Web can make use of Web Mining techniques. A truly semantic understanding of Web usage needs to take into account not only the information stored in server logs, but also the *meaning* that is constituted by the sets and sequences of Web page accesses. The examples provided show the potential benefits of further research in this integration attempt.

One important focus is to make search engines and other programs able to better understand the content of Web pages and sites. This is reflected in the wealth of research efforts that model pages in terms of an ontology of the content. Overall, three important directions for further interdisciplinary cooperation between mining and application experts in Semantic Web Usage Mining have been identified: (1) the development of ontologies of complex behaviour, (2) the deployment of these ontologies in Semantic Web description and mining tools, and (3) continued research into methods and tools that allow the integration of both experts' and users' background knowledge into the mining cycle.

Web mining methods should increasingly treat content, structure, and usage in an integrated fashion in iterated cycles of *extracting* and *utilizing* semantics, to be able to understand and (re)shape the Web.

Bibliography

[1] C.C. Aggarwal. Collaborative crawling: Mining user experiences for topical resource discovery. In *KDD - 2002 – Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, CA, July 23-26, 2002*, pages 423–428, New York, 2002. ACM.

[2] M. Baumgarten, A.G. Büchner, S.S. Anand, M.D. Mulvenna, and J.G. Hughes. User-driven navigation pattern discovery from internet data. In M. Spiliopoulou and B. Masand, editors, *Advances in Web Usage Analysis and User Profiling*, pages 74–91. Springer, Berlin, 2000.

[3] B. Berendt. Detail and context in web usage mining: Coarsening and visualizing sequences. In R. Kohavi, B.M. Masand, M. Spiliopoulou, and J. Srivastava, editors, *WEBKDD 2001 – Mining Web Log Data Across All Customer Touch Points*, pages 1–24. Springer-Verlag, Berlin Heidelberg, 2002b.

[4] B. Berendt and M. Spiliopoulou. Analysing navigation behaviour in web sites integrating multiple information systems. *The VLDB Journal*, 9(1):56–75, 2000.

[5] Bettina Berendt, Andreas Hotho, and Gerd Stumme. Towards semantic web mining. In *[20]*, pages 264–278.

[6] J. L. Borges and M. Levene. Data mining of user navigation patterns. In M. Spiliopoulou and B. Masand, editors, *Advances in Web Usage Analysis and User Profiling*, pages 92–111. Springer, Berlin, 2000.

[7] Erol Bozsak, Marc Ehrig, Siegfried Handschuh, Andreas Hotho, Alexander Maedche, Boris Motik, Daniel Oberle, Christoph Schmitz, Steffen Staab, Ljiljana Stojanovic, Nenad Stojanovic, Rudi Studer, Gerd Stumme, York Sure, Julien Tane, Raphael Volz, and Valentin Zacharias. Kaon - towards a large scale semantic web. In Kurt Bauknecht, A. Min Tjoa, and Gerald Quirchmayr, editors, *E-Commerce and Web Technologies, Third International Conference, EC-Web 2002, Aix-en-Provence, France, September 2-6, 2002, Proceedings*, volume 2455 of *Lecture Notes in Computer Science*, pages 304–313. Springer, 2002.

[8] B. Buchanan. Informed knowledge discovery: Using prior knowledge in discovery programs. In *KDD 2000 – Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Boston, MA, August 20-23, 2000*, page 3, New York, 2000. ACM.

[9] S. Chakrabarti. Data mining for hypertext: A tutorial survey. *SIGKDD Explorations*, 1:1–11, 2000.

[10] C. Chen. *Information Visualisation and Virtual Environments*. Springer, London, 1999.

[11] E.H. Chi, P. Pirolli, K. Chen, and J. Pitkow. Using information scent to model user information needs and actions on the web. In *Proceedings of the ACM CHI 2001 Conference on Human Factors in Computing Systems, Seattle, WA, March 31–April 4, 2001*, Amsterdam: ACM Press, 2001.

[12] E.H. Chi, P. Pirolli, and J. Pitkow. The scent of a site: a system for analyzing and predicting information scent, usage, and usability of a web site. In *Proceedings of the ACM CHI 2000 Conference on Human Factors in Computing Systems*, pages 161–168, Amsterdam: ACM Press., 2000.

[13] R. Cooley. *Web Usage Mining: Discovery and Application of Interesting Patterns from Web Data*. PhD thesis, University of Minnesota, Faculty of the Graduate School, 2000.

[14] R. Cooley, B. Mobasher, and J. Srivastava. Web mining: Information and pattern discovery on the world wide web. In *Proceedings of the Ninth IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97), November 1997*, 1997.

[15] M. Craven, D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam, and S. Slattery. Learning to construct knowledge bases from the world wide web. *Artificial Intelligence*, 118(1-2):69–113, 2000.

[16] H. Dai and B. Mobasher. Using ontologies to discover domain-level web usage profiles. In *Proceedings of the Second Semantic Web Mining Workshop at PKDD 2001, Helsinki, Finland, August 20, 2002*, 2002.

[17] Dieter Fensel, Stefan Decker, Michael Erdmann, and Rudi Studer. Ontobroker in a nutshell. In *European Conference on Digital Libraries*, pages 663–664, 1998.

[18] T. R. Gruber. Towards Principles for the Design of Ontologies Used for Knowledge Sharing. In N. Guarino and R. Poli, editors, *Formal Ontology in Conceptual Analysis and Knowledge Representation*, Deventer, The Netherlands, 1993. Kluwer Academic Publishers.

[19] Han and Kamber. *Data Mining. Concepts and Techniques*. Morgan Kaufmann, San Francisco, LA, 2001.

[20] Ian Horrocks and James A. Hendler, editors. *The Semantic Web - ISWC 2002, First International Semantic Web Conference, Sardinia, Italy, June 9-12, 2002, Proceedings*, volume 2342 of *Lecture Notes in Computer Science*. Springer, 2002.

[21] A. Hotho, A. Maedche, S. Staab, and R. Studer. SEAL-II — the soft spot between richly structured and unstructured knowledge. *Journal of Universal Computer Science (J.UCS)*, 7(7):566–590, 2001.

[22] J.Z. Huang, M. Ng, W.-K. Ching, J. Ng, and D. Cheung. A cube model and cluster analysis for web access sessions. In R. Kohavi, B.M. Masand, M. Spiliopoulou, and J. Srivastava, editors, *WEBKDD 2001 – Mining Web Log Data Across All Customer Touch Points*, pages 48–67. Springer-Verlag, Berlin Heidelberg, 2002b.

[23] T. Joachims. Optimizing search engines using clickthrough data. In *KDD - 2002 – Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, CA, July 23-26, 2002*, pages 133–142, New York, 2002. ACM.

[24] T. Joachims, D. Freitag, and T. Mitchell. Webwatcher: A tour guide for the world wide web. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 1997.

[25] H. Kato, T. Nakayama, and Y. Yamane. Navigation analysis tool based on the correlation between contents distribution and access patterns. In *Working Notes of the Workshop on Web Mining for E-Commerce - Challenges and Opportunities (WebKDD 2000) at the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 95–104, Boston, MA, 2000.

[26] C. Kemp and K. Ramamohanarao. Long-term learning for web search engines. In *Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD 2002), Helsinki, Finland, August 2002*, pages 263–274, Berlin – Heidelberg, 2002. Springer.

[27] Michael Kifer, Georg Lausen, and James Wu. Logical foundations of object-oriented and frame-based languages. *Journal of the ACM*, 42:741–843, 1995.

[28] R. Kimball and R. Merx. *The Data Webhouse Toolkit – Building Web-Enabled Data Warehouse*. Wiley Computer Publishing, New York, 2000.

[29] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[30] A.H.F. Laender, B.A. Ribeiro-Neto, A.S. da Silva, and J.S. Teixeira. A brief survey of web data extraction tools. *SIGMOD Record*, 31(2):84–93, 2002.

[31] W. Lin, S.A. Alvarez, and C. Ruiz. Efficient adaptive-support association rule mining for recommender systems. *Data Mining and Knowledge Discovery*, 6:83–105, 2002.

[32] A. Maedche and S. Staab. Discovering conceptual relations from text. In *ECAI-2000 - European Conference on Artificial Intelligence. Proceedings of the 13th European Conference on Artificial Intelligence*, pages 321–325. IOS Press, Amsterdam, 2000.

[33] P. Melville, R.J. Mooney, and R. Nagarajan. Content-boosted collaborative filtering. In *Proceedings of the ACM SIGIR Workshop on Recommender Systems, New Orleans, LA, September 2001*, 2001.

[34] D.A. Menascé, V. Almeida, R. Fonseca, and M.A. Mendes. A methodology for workload characterization of e-commerce sites. In *Proceedings of the ACM Conference on Electronic Commerce, Denver, CO, November 1999*, New York, 1999. ACM.

[35] B. Mobasher, R. Cooley, and J. Srivastava. Automatic personalization based on web usage mining. *Communications of the ACM*, 43(8):142–151, 2000.

[36] B. Mobasher, H. Dai, T. Luo, Y. Sun, and J. Zhu. Integrating web usage and content mining for more effective personalization. In *Proceedings of the International Conference on E-Commerce and Web Technologies (ECWeb2000)*, pages 165–176, Greenwich, UK, 2000.

[37] Resource Description Framework (RDF): Model and Syntax specification. W3C recommendation, 22 February 1999, www.w3.org/TR/REC-rdf-syntax-19990222/, 1999.

[38] W. Moe. Buying, searching, or browsing: Differentiating between online shoppers using in-store navigational clickstream. *Journal of Consumer Psychology*, (forthcoming), 2001.

[39] D. Oberle. *Semantic Community Web Portals - Personalization, Studienarbeit*. Universität Karlsruhe, 2000.

[40] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. In *Proceedings of the 7th International World Wide Web Conference*, pages 161–172, Brisbane, Australia, 1998.

[41] S. Parent, B. Mobasher, and S. Lytinen. An adaptive agent for web exploration based of concept hierarchies. In *Proceedings of the 9th International Conference on Human Computer Interaction*, New Orleans, LA, 2001.

[42] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.

[43] M. Spiliopoulou. The laborious way from data mining to web mining. *International Journal of Computer Systems, Science, & Engineering*, 14:113–126, 1999.

[44] M. Spiliopoulou and C. Pohle. Data mining for measuring and improving the success of web sites. *Data Mining and Knowledge Discovery*, 5:85–14, 2001.

[45] M. Spiliopoulou, Carsten Pohle, and Max Teltzrow. Modelling and mining web site usage strategies. In *Proceedings of the Multi-Konferenz Wirtschaftsinformatik, Nürnberg, Germany, Sept. 9-11, 2002*, 2002.

[46] R. Srikant and R. Agrawal. Mining generalized association rules. In *Proceedings of the 21st International Conference on Very Large Databases, Zurich, Switzerland, September 1995*, pages 407–419, 1995.

[47] J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan. Web usage mining: discovery and application of usage patterns from web data. *SIGKDD Explorations*, 1(2):12–23, 2000.

[48] Rudi Studer, V. Richard Benjamins, and Dieter Fensel. Knowledge engineering: Principles and methods. *Data Knowledge Engineering*, 25(1-2):161–197, 1998.

[49] G. Stumme. Conceptual on-line analytical processing. In K. Tanaka, S. Ghandeharizadeh, and Y. Kambayashi, editors, *Information Organization and Databases*, chapter 14, pages 191–203. Boston–Dordrecht–London, 2002.

[50] Gerd Stumme. Using ontologies and formal concept analysis for organizing business knowledge. In *Proc. Referenzmodellierung 2001 (in print)*, 2002.

[51] Y. Sure, M. Erdmann, J. Angele, S. Staab, R. Studer, and D. Wenke. OntoEdit: Collaborative ontology development for the semantic web. In *[20]*, pages 221–235.

[52] Y. Sure, S. Staab, and J. Angele. OntoEdit: Guiding ontology development by methodology and inferencing. In *Proceedings of the International Conference on Ontologies, Databases and Applications of SEmantics ODBASE 2002*, University of California, Irvine, USA, 2002.

[53] A. Ypma and T. Heskes. Categorization of web pages and user clustering with mixtures of hidden markov models. In *Workshop Notes of Fourth WEBKDD Web Mining for Usage Patterns & User Profiles at KDD - 2002, July 23, 2002*, pages 31–43, 2002.

[54] O.R. Zaiane, M. Xin, and J. Han. Discovering web access patterns and trends by applying olap and data mining technology on web logs. In *Proceedings of Advances in Digital Libraries Conference (ADL'98), Santa Barbara, CA, April 1998*, pages 19–29, 1998.