Electronic market: The roadmap for university libraries and members to survive in the information jungle

Michael Christoffel, Sebastian Pulkowski. Bethina Schmitt, Peter C. Lockemann

Institute for Program Structures and Data Organization
University of Karlsruhe
D-76128, Karlsruhe, GERMANY
E-mail: {christof, pulkowsk, schmitt, lockeman}@ira.uka.de

Abstract

This contribution argues that electronic markets can serve as a powerful mechanism to entice providers to identify their customer base and to offer customer-oriented, high-quality and economical services and to induce customers to a more focused and price-conscious behavior. The paper claims that this should be particularly true for the provision and access to scientific literature where the tradition so far has been mostly free access by customers and non-transparent cost accounting and service procurement by university libraries. We report on a project for developing a technical network infrastructure that allows for a more cost-transparent access to scientific literature by campus users and attempts to add a competitive element to library services. Equally important, it provides added value to the users so that they can orient themselves in the vast expanses of scientific literature much faster and more economically. We cover three major elements of the infrastructure: user agents, traders and source wrappers.

1 Introduction

1.1 Digital Libraries as an open market

Access to scientific literature is currently undergoing drastic changes. This is due to open system network facilities and the many new activities that try to exploit them in order to make documents available in electronic form. The documents range from journal articles and conference papers all the way to dissertations, technical reports and entire books. These changes have a particularly strong impact on universities, where traditionally there has been a continuous and heavy demand on literature supply. On the one hand, the campus can be expected to profit from the much improved opportunities for literature search and acquisition. On the other hand, it experiences a number of ever more serious problems, both from a provider and a customer perspective.

From the perspective of the provider, the market may become one of fierce competition, where traditional literature providers are threatened by novel and attractive service offerings. This seems particularly true for university libraries which traditionally have held a near monopoly on literature provision on campus. In Germany we are observing a wealth of new services such as electronic document delivery, electronic book ordering, electronic journal subscription and delivery, literature search and profiling services. This is over and above the more traditional library services of catalogue search, book loaning and reading rooms, and make even the latter services ubiquitous, i.e. no longer physically bound to the nearest library. In fact, university libraries start to rise to the challenge by collaborating on the new electronic offerings. Nonetheless, other institutions - e.g., public libraries, commercial search services, publishers - may join with similar offerings so that the user will have a choice between a number of comparable, conveniently via WWW accessible services, be in a position to compare them according to his own criteria, and finally decide on the most attractive offer. The situation is rapidly moving to something which looks like an open market of document services, where the laws of supply and demand hold and only the most competitive providers survive.

University libraries will have to learn to operate within such a market, to identify their particular strengths and competencies, to value them, and to be able to charge a price. We claim that university libraries have indeed such strengths due to a large base of local customers whose interests they know especially well, have ready access to and are easily adjusted to. As such they can provide tailored search facilities and can negotiate special contracts with off-campus document providers for campus licenses at attractive prices. Since many university libraries will pursue similar objectives, one will also have to expect a competitive market among them, where only a few will survive the competition.

A customer seems to draw a number of benefits from the new situation. He can operate on a much larger document space, the search can be much more focused, and the noteworthy documents can be retrieved within much shorter time and even in a variety of formats at a variety of prices. Unfortunately, that is true only in principle.

In reality, the customer experiences a number of obstacles. Operating on ever growing document supplies results in what is called information overload. A focused search requires the capability to articulate one's needs in sufficiently precise terms, which seems only possible if there exists considerable knowledge on the existence of the available services and their functioning. In order to determine the optimal delivery circumstances one must analyze and compare the conditions of the various providers, such as delivery time and form - a tedious and time-consuming affair.

Until now, campus users have generally been accustomed to free library services. Hence, they were rarely forced to attach a monetary value to their requests. This gave rise to retrieval strategies of a roaming nature: Searches start with an at best approximate description, and the user then iterates through a number of steps that narrow or shift the focus depending on the results of the previous steps. In a market where every service rendered has its price, this behavior is bound to change. Since the essence of scientific work is being knowledgeable of the current state-of-the-art, all our efforts should go into the direction of the campus user availing himself of document search and retrieval possibilities under the new circumstances, rather than being repelled by it. He must be put in a position where he can, right from the beginning, give a sufficiently precise formulation of the objectives; not only in terms of document contents but also of time, price, and additional constraints.

1.2 The UniCats Project

The objectives of our project[1] are twofold: to develop - under the premise of an open market for document services - a technical infrastructure that

- allows campus users, under their individually defined, optimal conditions, to locate and retrieve documents from a network, transparently if so desired
- enables university libraries to compete effectively in the market for a base of academic-oriented customers.

The infrastructure is based on a coarse architecture that encompasses three components:

- User agents should offer the user appropriate capabilities which allow him to interact in a uniform fashion with a complex system, and to counteract information overload.
- Traders administer metadata concerning existing services, in order to be able to localize the most favorable source in fulfilling a query. This isolates the user from the knowledge of the existence and functionality of available services.
- Wrappers hide the syntactical and semantic heterogeneity of individual services. They can therefore be addressed within the network in a uniform and possibly parallel manner.

Above and beyond this, cost issues and payment systems are an essential part of the architecture. Through the interaction of the above-mentioned three components, conditions for a market economy situation are created, which is easy to comprehend and thus easy to use. This is the motive behind our project name UniCats: "a UNiversal Integration of Catalogues based on an Agent-supported Trading and wrapping System".

With these intentions in mind, we are pursuing in the first phase the goal to create the necessary infrastructure in order to be able to experiment with pricing models. It is only in this manner that one can gain the feel of the value of information to the user, thereby influencing the fee structure of the providers.

We are limiting our scenario to the University of Karlsruhe, and would like to corroborate our thesis, that university libraries can clearly remain competitive and are likewise able to offer attractive services to the campus. In the future, even libraries will need to charge for their services, due to the overall shortage of funding. This project provides a good experimental field for various pricing models and student support models. This is the only method for the client as well as the provider of making clear the actual worth of the various needs for information, various research services and various delivery services.

Challenges and concepts for solutions of the three UniCats components will be outlined in detail in the following sections. We will conclude with an account on the existing status and the planned developments of the project.

2 The User Agent

The user agent is to represent the interests of the client. This means offering a uniform, comprehensible, and user-friendly interface to the diverse services related to literature research, along with a mechanism to protect the client from a flood of information. The user agent is the interface between user and system such that the client can profit from the various services and influence the market according to his needs.

The interaction of the three components of the UniCats architecture, as seen from the viewpoint of the user agent, is as follows: The client formulates a query. In order to fulfil this query, the user agent requires the assistance of the two other UniCats components. This way he can consult the trader for the appropriate providers. Following this, the user agent turns to the selected wrappers with a uniform query format, collects the separate results and integrates them. The final result will then be presented to the user in an appropriate display.

Concepts used for query support, as well as strategy alternatives towards the fulfillment of a query, and the integration and presentation of results is detailed below.

2.1 Formulation of a Query

A simple keyword search query is the most widely used method in literature research. By using this method, the user is flooded with information using even a middle-sized catalogue. An integrated system such as ours, providing much more information than a single catalogue, especially requires mechanisms which counteract this flood of information.

[1]The project is supported by the German Research Foudation (DFG)

The more explicit a user can specify his query, the more precise is the information to be delivered. Our system not only handles descriptions of content, but also takes delivery parameters, such as format, time, and costs into consideration. These are to be included in the formulation of the query. Nevertheless, this does not correlate with the desire of the user, who would like to begin researching as soon as possible with the least amount of initial input.

Our proposal to simplify the formulation of the query is a role model for the university environment: the user, as before, is to begin his search with a keyword, but must then additionally assume a predefined role. For instance, a "student" will be recommended a few textbooks in his native language, which are relatively new and immediately available. A "young researcher" obtains in addition to advanced textbooks the current periodicals and conference articles, with an emphasis on surveys that summarize the current state of the art. Also important is a role of "impatient", in which only those documents are displayed which are immediately available, even at considerable cost.

We expect that a suitable role constellation will evolve only after a certain a longer period of experimentation. For the first prototype in our project scenario we shall employ the aforementioned roles. In the longer run, we look for more general facilities, through which users can formulate their needs for information and the importance they attach to them. This will give rise to a more refined role model.

2.2 Evaluation Strategies

After the client has formulated his query, the user agent must develop a strategy or protocol, i.e. decide which service in which order, possibly simultaneously, he is to address. The user agent has thereby a wide spectrum of possibilities: it can, for instance, address that source which the user explicitly defined, or it can address those sources which the trader recommends for a query. In the first case, the user agent surrenders control to the user. Since the user should in general be relieved of this decision, the option should only be made available on request and to expert users. In the second case, the user agent passes the decision on to the most competent location in the system: the trader.

In a market economy a trader can be expected to demand payment for its service so that the user agent should principally be able to act without this help. In fact, this should become possible if the user agent learns from past requests, or has sufficient knowledge of the user's preferences. The user agent collects information on individual users as well as on the entire group of users, so that this information is available to it when taking a decision.

2.3 Integration of Results

In order to counteract information overload, and to spare the user the time and effort of comparing documents in result collections from different sources, the results should automatically be agglomerated. To do this, identical documents in separate result sets must be identified and the available information brought together. Availability times for different libraries, the retail price for an online order, as well as a copy of the cover, the table of contents, a summary, or even a review can be displayed for one document.

Existing identification schemes for library documentation are not necessarily reliable (ISBN), nor universally accepted (LCCN, CODEN), nor offering sufficient depth (ISSN). Duplicates cannot reliably be recognized through the comparison of these singular attributes.

Therefore, as a rule, recognition of duplicates will take place through duplicate-control-numbers or match codes, using various attribute combinations and a number of information-extracting methods (ABC, USBC). Title, author, place of publication, year of publication, edition, ISBN and ISSN, are used most often for this purpose. Certain parts will be extracted from these fields following normalization (elimination of spaces, symbols, and capitalization).

In our scenario, the collection of results arises incrementally; resulting documents reveal only an abbreviated entry consisting of title, author, and year. Thus we must do without the usual preparatory phase of clustering potential duplicates, and follow through with recognizing duplicates in terms of their abbreviated entries.

2.4 Presenting Results

The manner in which results are presented is of significant consequence to the acceptance of any interactive system. One can even counteract the flood of information through clever visual graphics. Nevertheless, speedy replies are demanded of every interactive system, which can be achieved, for example, through text-only displays. We believe that in general there is so much interesting information for a document that better recognition is achieved by presentation in the form of graphics. Just take the non-contents information on accessibility, or the type of document. A graphic overview clarifies the composition of the resulting set in a very concise manner. Thus one can visualize this information through abstract forms, or in the case of our scenario as concrete metaphors, as these considerably ease access by the novice user.

In order to clarify the connection between concept and document space for the user, post-processing operators are necessary for the results [Wan97]. Only where the user has a clear grasp of the composition of the results can he judge whether or not his search concept was properly selected or in which manner he has to correct it to obtain the desired result. Essentially, one needs operators to sort and group, as well as extend or reduce the resulting set. The referred documents are independent of various views that can be used to display them [Kra88]. Thus documents can not only be displayed textually as a hit list, in the form of tables, or as Hi-Cites [Mic98], but also as tangible 3-D books sorted onto shelves.

Again by experimentation we hope, through observation of active use, to be able to determine which displays are most appropriate, and therefore preferred by most literature researchers. There is no means to predict this in advance.

3 Trader

Traders have the task of mediating information providers to a customer (or a user agent) looking for information, which will give him the greatest advantage in the research and delivery of the desired informations and documents [Bea97, Dre98]. In addition to the address of the wrapper assigned to the information source, the trader transmits informations about the provider, such as the available types of document delivery, focus of the offered documents and services and an estimate of incurring costs. The research and delivery services on documents are not the traders responsibility, traders strictly restrict themselves to the mediation service. The trader receives a rating of mediated providers as a response from the user agent. Expense calculation and planning are of primary concern for a competitive trader.

To be able to fulfil its task, the trader must have a thorough knowledge of the market development. It stands in connection to user agents, wrappers and other traders. We plan to take a closer look at the cooperation between the trader and these components, as well as his fundamental operating methods.

3.1 Characterization of the Information Providers

The trader keeps characterizations of those information providers which it is to mediate, in the form of descriptive attributes. These attributes describe, among other, the type of the provider (e.g. library, research service, specialist publisher), offered thematic areas, delivery times, charging models. The profile of the individual provider is a set of assignments of values to these attributes. This profile may characterize the library of the department of art history as well as the full text delivery service for technical reports concerning computer science.

While linking a provider to the UniCats system, his profile will be created by a person with a good knowledge of the source, such as a librarian. Should the characteristic of the source change later on, the profile can be updated at any time.

By means of test queries, errors and omissions in the profile can be eliminated, and dependencies between the attributes detected. Moreover, it is possible to determine values of attributes, which cannot be assigned while linking a source, in an experimental way during operation, for example the average response time of the source.

3.2 Query Handling

In addition to the precise knowledge of the providers, precise knowledge of the customer's needs is necessary. Therefore, a description of the desired information and the desired information provider is handed over by the user agent in his query. Besides the attributes declared by the providers, attributes, which access the experience of the trader, may be specified in the user query, such as an upper time limit or a maximum price for the given research. In order to weigh the attributes, a fuzzy-value can be attached to any value of the attributes.

To clarify the customers demand, an approach based on neighborhoods in the attribute space is used. The neighborhoods are computed from similarity relations over the attributes, which are expressed as values within the interval [0,1]. This allows to express general knowledge about the attributes: "Geometry is an area of mathematics" or "a delivery time of one hour is better than a delivery time of two hours" may be examples.

The trader examines the information providers known to it, determines the relevant ones and sorts them according to the conformity with the customers demand. The degree of conformity between a profile and a user query can be described by a mathematical function and calculated for concrete input data. The trader also considers the restrictions specified in the query for the choice of the provider, such as the limitation to free or local providers. We foresee some learning capability for the trader to use its experiences for the choice and evaluation of the attached sources.

Experiences are gained experimentally through test queries, and questionnaire-like through the use of ratings which the customer (or the user agent) returns. This feedback is essential to the observation and analysis of the market development. In addition to the direct rating of every mediated provider through a fuzzy-value, the trader receives further information needed for a statistical analysis. This includes, for example, the total time for the research, the number of located documents of value to the customer and the actual accumulated cost.

3.3 Trader Federation

For a practical, attractive and scalable environment, the presence of more than one trader is necessary. However, the customer should not be confronted with a new selection problem, but receive further assistance by the cooperation of the traders. Hence, the traders join into a trader federation. This federation is organized and managed by its members on their own. For this, centralized and decentralized models are presently being studied.

The federation is hierarchically organized in the form of a tree, so that a trader has any amount of subordinate traders, but at most one superior trader. This structure opens the possibility for every trader to specialize in a field and to choose the providers mediated by it. For instance, a trader could specialize in documents concerning engineering; then this trader could have a subordinate trader specialized in electrical engineering. Every trader determines its own profile as the union of the profiles of the providers directly connected to it and the profiles of the subordinate traders.

The traders joined in the federation support each other in handling those queries directed towards them: If in the query handling a high conformity between the user query and the profile of a subordinate trader is found, then the query will be forwarded to this trader. A query is forwarded to the superior trader only if the minimal number of providers desired by the customer could not be found. This strategy precludes a query from being forwarded simultaneously to every trader in the federation.

The customer's ranking is always passed on to all traders

that were involved in the federated reply. This includes the originally addressed trader that gains valuable insights into the services rendered by the other traders.

If no providers, or not enough providers, are found in the trader federation while handling a user query, then the trader will simplify the query by reducing the number of attributes. The evaluation process will be restarted using this new query.

3.4 Charges

For providers who charge, charging and accounting systems must be integrated in the trader. In addition, commercial traders have their own charging models. The actual charge may be both billed to the customer for the mediation of the sources and to the provider, since he has an advantage from being mediated by the trader; naturally, the charge may be split between the customer and the provider.

The costs of an individual trader must be taken into consideration when handling a query. Before forwarding a query to another trader, the necessary costs must be calculated in advance, so that the trader can decide whether or not there is sufficient value for the additional costs.

Various charging models are being studied and will be tested in a working environment after completion of the system. One is a performance-oriented model in which the incurred payment depends on the number of the information providers that were mediated to. Another extends this model by measuring the benefit the user gains through the mediated providers.

However, a performance-oriented model may not be practical to the provider, because he does not possess an overview of the customers he has been mediated to. A time based flat rate model is more suitable: A flat rate is set when registering a provider for a period of time. After expiration of this period, the registration can be extended, or the provider can relinquish the trader and connect to another.

4 Wrapper

A wrapper is a kind of translator, which takes queries (in this case those of the user agent), and syntactically and semantically reshapes them so that they may be processed by the source. Following this, the results will be recomposed to a response that the user understands. Since almost all information-providers now offer an HTML interface for their services, we will initially concentrate on wrapping HTML sources.

4.1 Existing Situation

Many projects [Ade98, Ash97. Fau98, Ham97] have concentrated on wrapper-construction for HTML pages in the last few years. Nevertheless, most of these wrappers are not or hardly useable in our scenario. All of them are exclusively generated from the source and limited to this. Further, the task of constructing a wrapper had to be given to a specialist who was required to adapt the wrapper to an individual information source. The reuse of a wrapper for another information source was not possible. Besides, most wrappers possess their own query and response formats, so that a UniCats user agent is unable to communicate with it. Another weakness from our point of view is that wrappers concentrate exclusively on the translation of query and response. The information sources in our scenario reveal peculiarities: one must in the course of a research visit many linked HTML pages from which individual bits of information are extracted and composed to create the end result. In other words, several pages with varying structures are visited by navigating through the source. Further, due to the trader connections meta-information concerning the source such as delivery and cost factors must be procured. Due to cost considerations, information access should be optimized. None of these can be handled by existing wrappers for HTML sources.

The complexity of wrappers does not lend itself to a simple construction. Therefore, administrators should be supported by suitable tools which we collectively refer to as a wrapper-generator. The possibility of easing wrapper creation is especially interesting when - this is one of our more interesting hypotheses - user organisations such as libraries create their own wrappers if a source is unable or unwilling to do so.

4.2 The UniCats Project Wrapper

We are developing a first wrapper-generator for the UniCats project that is easy to be used by both the source administrator and the user organisation, and at the same time meets their needs. The approach can be dubbed "generation by example": The administrator who generates the wrapper carries out an investigation in the source to be wrapped. During his sample investigation, he specifies both the content and structure of a page. This is to be done with all the pages required for later access by the user agent. As much meta-information as possible is to be recorded parallel to inputing the specification. These encompass attributes and result formats of the document required for inquiring the wrapper. Meta-data should be invested with additional source-information to be transferred to the trader, so that this latter may mediate the source.

Especially in the area of digital libraries, sources differ widely in function and content. Hence, a modular concept is needed to compose a wrapper individually for a given source. Several modules have been developed which cover, for example, areas of cost surveillance, completion of the questionnaire, or the conversion of resulting pages. The modules can be assembled individually for a wrapper so that it possesses only those functions supported by the source. In addition, various strategies can be chosen in order to influence the sequence of page accesses and, hence, the total cost of the inquiry. Often there is more than one possibility to reach a page that finally displays the desired information. It is thus quite conceivable that two wrappers for the same source may differ in execution time and cost. This adds another element of competition if wrappers are provided by competing administrations.

4.3 Conflicts of Interest

The possibility that a wrapper for a given source may be supplied by more than one administration, and especially

if one is the source administration and the other the user organization will give rise to conflicts of interest. Clearly, no such conflicts exist as long as an information source does not have any financial interest, as in the case of the technical reports of the universities. Otherwise, a wrapper written by the source provider will presumably optimize the profit for the source. The counteracting influence is user satisfisfaction with regard to service and price. Should the user have the feeling he has paid too much for the service, he will try to find another wrapper for this source. He may even try to generate his own and then obviously try to optimize his own cost.

Conflict can arise in connection with cost control. A wrapper provided by a source has all the access rights to determine and convey the actual costs. As opposed to this, the wrapper of a user organisation will probably never have support by the given source in obtaining this information. The wrapper can at best try to estimate the costs of research in a source on the basis of mostly imprecise information.

Another problem arises, when a source provides information, either false or overly exaggerated statements concerning capabilities, costs, and source content. A user or user agent must first recognize such a situation. Punishment may follow, and could, for instance, be administered in the form of exclusion by the trader. It is expected that such problems will be dealt by virtue of competition alone.

The scenario can be extended to wrappers provided by a so-called third party, which can be used by a user at cost. It is now this wrapper that must find a compromise between the interests of user and source. This may have to be considered in the creation of a wrapper, and the strategies towards information collection.

5 Conclusion

Presently, each of the components described as well as the necessary communication structure between these components are being implemented in a first version in Java. A prototype is being planned for the beginning of next year.

In the area of accounting systems, additional developments and innovations are expected to bring about suitable methods for use in our scenario. In the beginning, experiments using fictional budgets can be applied, or payment will continue to be handled by the already existing user accounts at the university library. The first studies on pricing models are to be made within the setting of our project. The actual value of services rendered must first be determined; meaning the time and money a user is willing to concede. We will also have to observe how campus users will adapt their methods of research, keeping in mind that in future costs will play an essential role. Until now, there exists no study regarding the research methods of a student under cost restraints. Initial experiences with various pricing models are to be gained in conjunction with our university library, in order to assess appropriate fees for individual services. This would be especially important for the university library, in order to learn about suitable marketing strategies of the future and survive in an open market.

References

[Ade98] Brad Adelberg. Nodose: A tool for semi-automatically extracting semi-structured data from text documents. In *SIGMOD Conference 1998*, pages 283–294, 1998.

[Ash97] Naveen Ashish and Craig Knoblock. Wrapper Generation for Semi-structured Internet Sources. *SIGMOD Record*, 26(4):8–15, 1997.

[Bea97] M. Bearman. Tutorial on ODP Trading Function. Technical report, Faculty of Information Science & Engineering, University of Canberra, Australia, Feb 1997.

[Dre98] M. Dregger, N. Fuhr et al. Provider Selection - Design and Implementaion of the Medoc Broker. In M. Breu et al. A. Barth, editor, *Digital Libraries in Computer Science: The MeDoc Approach*, Lecture Notes in Computer Science, pages 67–68. Springer, Heidelberg, Germany, 1998.

[Fau98] Lukas C. Faulstich and Myra Spiliopoulou. Building HyperNavigation wrappers for publisher websites. In *Second European Conference on Digital Libraries*, Heraklion, Greece, September 1998.

[Ham97] J. Hammer, H. Garcia-Molina, J. Cho, A. Crespo, R. Aranha. Extracting Semistructured Information from the Web. In *Proceedings of the Workshop on Management of Semistructured Data*, pages 18–25, Tucson, Arizona, May 1997.

[Kra88] G. E. Krasner and S. T. Pope. A Cookbook for Using the Model-View-Controller User Interface Paradigm in Smalltalk-80. *Journal of Object-Oriented Programming*, 1(3):26–49, August 1988.

[Mic98] Michelle Q Wang Baldonado and Terry Winograd. Hi-cites: dynamically created citations with active highlighting. In *CHI '98. Conference proceedings on Human factors in computing systems*, pages 408–415, Los Angeles, CA, April 18-23 1998.

[Wan97] Michelle Q Wang Baldonado and Terry Winograd. SenseMaker: An information-exploration interface supporting the contextual evolution of a user's interests. In *Proceedings of the Conference on Human Factors in Computing Systems, CHI'97*, pages 11–18, Atlanta, Ga., March 1997. ACM Press, New York.