Concepts for an Intelligent Information Portal in Pharmaceutical Research

Alex Kohn, François Bry
(Inst. Inf., University of Munich, Germany)

Stefan Klostermann, Alexander Manta
(Roche Diagnostics GmbH, Penzberg, Germany)

Abstract: Once upon a time scientists were experts in their field. They knew not only the "hot questions" but also the scientists involved and the various approaches investigated. More important, they were well informed of novel research results. Gone are these favorable times! Hot issues and active research teams emerge with high pace and being informed within days or even hours might be essential for success. Furthermore, no one can any longer keep an eye on the research publications, patents, and other information that might be relevant for one's research. As a consequence, scientists often feel - and in fact they sometimes are - rather unaware of areas that are of prime importance for their research. High diversity, considerable amounts of information and extremely fast communication are key characteristics of today's research - especially in medical biology. An automatic tracking of technical and scientific information is a way to cope with these aspects of today's research. Such a system is made possible by emerging techniques such as "Semantic Web". This article describes the corner stones of such an "Intelligent Information Portal" currently developed at Roche Diagnostics GmbH for scientists in Pharmaceutical Research. The article stresses the salient aspects of the envisioned system and its focus on personalization/adaptation.
Key Words: adaptation, data integration, knowledge management, life sciences
Category: I.2.0, I.2.1, H.3.0, H.5.0

1 Introduction

Since the beginning of the information age, the data output has steadily increased. Hot research topics emerge with such a high pace that being informed fast might be eminent for success. The mere number of 1700 new publications appearing daily on Medline[1] only (in 2006), shows how dramatic the information overload is that a scientist has to deal with nowadays. Extrapolating to other domains like patent or sequence databases worsens the problem. Obviously, nobody can continuously keep track of all the information being published that might be relevant for one's research. The quantities are too high, changes too fast and data too heterogeneous. The situation observed in the public domain has analogies in many corporate intranet environments, as it is the case for instance at Roche Diagnostics GmbH.

The present problems of information retrieval in a Pharma Research department will be outlined by a scenario. Let's assume that a scientist has identified

[1] After the key Medline indicators: http://www.nlm.nih.gov/bsd/bsd_key.html

an oncology target (ReceptorX) expressed in human breast cancer cells, which is a target for therapeutic proteins (e.g. antibodies). Inhibiting/activating the receptor results in a modification of the signaling pathway. The consequence is apoptosis of breast cancer cells expressing ReceptorX. Prior to clinical studies in humans, model organisms are used to test the biological effects of the potential drug. The identification of suitable models requires a homology search of ReceptorX in different species. A toxicological model is only suitable if the homolog is recognized by the therapeutic antibody in question. Answering the scientist's query, which could be phrased as "homologs ReceptorX", involves several steps:

Fundamentals: The starting point for the query is an in house gene database. Here, fundamental information like gene structure, synonyms, some homologs, literature, etc. is available.

Homologs: The homologs returned by the gene database will usually not suffice and therefore a homology search against other sequence databases (species) is necessary. For that task a couple of web-based sequence analysis tools (e.g. BLAST[2]) are available to the scientist.

Experts: Sequence analysis tools have many parameters and false settings can lead to poor results. Assuring quality requires finding out who is an expert in this area. Furthermore, the expert's contact details need to be acquired by looking into the company's telephone book.

Candidates: Having found all homologs, the scientist would like to know if any of them is currently part of a project. Hence, a query against corporate databases, tracking all projects and lab experiments in protein research has to be performed.

Reflecting on this scenario, several issues arise: S/he must know which resources to use, s/he must know where to find them and s/he can't query all four resources at once. These observations can be generalized, such that the characteristics of the information landscape we face, get apparent. We have many *heterogeneous resources* like databases, applications, websites, web portals (SharePoint, LiveLink, etc.) and file shares. The resources generally have a *low linkage*, hence *low visibility*: users are often not aware of the existence of relevant information resources. However, one of the biggest shortcomings is the *lack of meta information* on these resources. All that makes information retrieval quite difficult [Mühlbacher 08].

Traditional intranet search engines and data integration approaches fail to cope with these issues, hence they fail to fulfill the information needs of a scientist. Intranet search engines perform poor [Xue 03] due to several reasons: the linkage structure is hierarchical thus the ranking-assumption of popularity equaling relevance can't be applied, access barriers (security) are common and

[2] Basic Local Alignment Search Tool: http://www.ncbi.nlm.nih.gov/BLAST/

the "deep intranet" is usually not indexed. In data integration, a traditional approach is data warehousing. Here several databases are integrated by replication into a central store. Besides its merits, this approach has also drawbacks: data is mirrored and thus out of date or can't be mirrored at all, IT costs are high, etc.

How would a perfect information landscape look like? It would consist of a system S which amongst others would have the following features: it has a user-centric interface, it can be easily used for search & navigation, it knows all the resources and their semantics, it has domain knowledge, it learns from its users, it is context-aware, it knows the users' information needs and last but not least it has more knowledge about itself than the user. S is a vision. Although not feasible in the near future, we could at least begin to move into this direction. A first step towards S will be the development of an Intelligent Information Portal (IIP), whose corner stones are described in the following sections.

2 Intelligent Information Portal

In light of S we identified five key cornerstones of an IIP: Resource Integration, Semantics, Ontology Management, Querying and Adaptation.

The resources (databases, applications, websites, web portals, file shares, etc.) need to be integrated and made accessible via a standardized interface. Many concepts for integrating resources exist. An example is the Kleisli [Davidson 97] query system developed at the University of Pennsylvania and marketed today by several companies (see [Section 3]). Another example is the concept of Service-Oriented Architecture (SOA). The paradigm of SOA is to distribute the system logic over several independent services rather than building a monolith system.

Semantic technologies enable the computer systems to understand and reason on data. The semantic web standards RDF (Resource Description Framework) and OWL (Web Ontology Language) provide a framework for adding meta information to both, the resource structure and the data itself. In the public domain many ontologies are already available, especially in biology which is a driving force. Alone on the Open Biomedical Ontology Foundry[3] website a huge collection of publicly available ontologies can be found. Amongst others, the well known GeneOntology[4] is available. To illustrate the doors semantic technologies can open, consider e.g. two databases DB_1 and DB_2 both storing taxonomical data. In DB_1 data is stored in a field called "species" while in DB_2 the field is called "organisms". By applying an ontology mapping to each database schema, the computer will infer that both fields in fact refer to the same concept. Besides the annotation of data and its structure, one could also provide meta information to the resource itself, i.e.: a description of what kind of data is available,

[3] http://obofoundry.org/
[4] http://www.geneontology.org/

where a resource is located, how a resource can be accessed, etc. Given this meta information and the mapping, the resource becomes a self-described modular data resource. Hence, by wrapping semantics around the existing systems, previously unconnected resources become related and a cross-domain query becomes possible. It has to be noticed, that a high-scale usage of ontologies requires an Ontology Management System, which has to address several curation tasks: it must define methods for storage, versioning, up-to-dateness, mediation, etc.

The next two subsections will describe adaptation and an advanced keyword query approach in more detail as these are the key components the user interacts with.

2.1 Adaptive Personalization

Provided that integration succeeds, a huge amount of information will be available at the researcher's desk. Obviously the risk of information overload remains, leading to poor precision and recall when doing inquiries. A promising technique to mitigate these issues is adaptive behavior. An *adaptive system* is a system which adapts its communication patterns to the current actor. *Recommender systems* are a specific implementation of this technique. They guide the user in a personalized way through the complex information landscape with its large number of options [Burke 02]. *Personalization* in this context means to adapt the communication pattern to the user's characteristics [Baldoni 05]. An essential part of adaptive systems are therefore *user profiles* which store user preferences in attribute-value pairs. The data stored in a user profile can contain amongst others [Baldoni 05]: device information, preferred settings, goal, current task, information needs, required information depth, time constraints, previously regarded information, previously gained knowledge and much more. The maintenance of a user profile, i.e. the acquisition of data, the update and inconsistency checking is accomplished by a *user model*.

A detailed description of recommendation techniques is given in [Burke 02] who distinguishes between five basic approaches, namely Collaborative-based, Content-based, Demographic-based, Utility-based and Knowledge-based recommendation. Relying on these basic approaches, several hybrid-based systems have been proposed. In order to give an idea of how recommender systems work, the collaborative filtering (CF) and demographic-based approaches are described briefly.

User or item-based CF is the most mature technique. It is used today by many applications especially in e-commerce (e.g. Amazon or E-Bay). The basic idea of user-based CF is to recommend previously unknown items to a user based on the items preferences in his neighborhood. Let U be the set of users, I the set of items and r_u a rating vector with $u \in U$, mapping items to a value of unity if it is considered relevant by u and zero otherwise. Given users

$u_1, u_2, u_3 \in U$, items $A, B, C \in I$, the ratings $r_{u_1} = \{(A,1),(B,1)\}$, $r_{u_2} = \{(C,1)\}$ and $r_{u_3} = \{(A,1),(B,1),(C,1)\}$. A significant correlation between u_1 and u_3 can be detected since both have rated item A and B positive. Thus, u_3 is in the neighborhood of u_1 and the unknown item C can be recommended to u_1. The idea of item-based CF is very similar to user-based CF. In item-based CF the perspective is opposite, meaning that highly correlated items are found according to the preferred items. Commonly used techniques in CF are Pearson correlation, vector similarity, clustering, Bayesian networks, etc.

Demographic recommenders classify users into classes based on their personal attributes. Therefore, it is eminent that users provide explicit personal information about their preferences. The information for categorization can be gathered with surveys or by the usage of machine learning algorithms which analyze user profiles. Given the demographic data, a system can identify demographically similar users to extrapolate for instance from their ratings.

Personalization relies on user profiles so that privacy issues arise. The following policies describe options of how to abate them. Most importantly, the works council and the users have to be elucidated about the stored data. Keeping the profiles transparent is also crucial. Users should have the possibility to view their profiles and eventually delete data that they don't want to be stored. A third approach is anonymization. This could be achieved by applying personalization on the group level instead of the individual. Here, roles, tasks, projects, etc. are pre-defined and the user can select between them in a multiple-choice manner. Depending on the groups a user has subscribed to, the information portal is adjusted. Given that a set of users belong to a common group, their actions will contribute to changes of the group profile. Thus in group recommenders, the system tries to fulfill the needs of all group members by maximizing the average member-satisfaction. While this approach guarantees anonymity, it has several drawbacks: (a) the initial choice of the proper groups is difficult, (b) group members drifting away with their interests will gain a poor personalization and (c) personalization can't be as accurate as applied on an individual level.

The pilot will apply only group-based recommendation as default. However, people will have the freedom to decide if they want an individual personalization activated or not. In both cases, the personalization will be transparent. Hence, inference rules will be shown, profile data is viewable and in case of individual personalization also erasable.

2.2 Advanced Keyword Query

A search in IIP will be keyword-based with an easy to use structural and/or semantic prefix extension, such that scientists are able to specify what they are interested in. On the one hand, the more detailed a query language is, the more accurate the delivered answers are. On the other hand, formal query languages

have to be learned and are therefore not readily used. Simple keyword-based query interfaces (e.g. Google or Yahoo) have without a doubt a reason for their success: little is required to enter a few keywords. However, keyword search, how simple and therefore appealing it might be, is unspecific. Looking e.g. for ReceptorX, one cannot distinguish between articles that accidentally mention the concept somewhere, and those that have the concept in their *main title*, *section titles*, etc. Thus, refining keyword query with some structural elements while keeping the appealing simplicity of keyword querying has been proposed by [Cohen 03].

Structural information can be added to a query by providing the user with a small number of structural concepts. This can be done with a simple textual query interface, e.g. "mt:ReceptorX" expressing that ReceptorX should occur in the *main title* or "t:ReceptorX" meaning that ReceptorX should occur in a *title* at any depth. Dependencies between components can be expressed as well, e.g. "s:ReceptorX > a:John Q. Public" expressing that a *section* contains ReceptorX authored by John Q. Public is sought for.

Accordingly, we propose semantic prefixes to be added to keyword-based query. Instead of using keyword prefixes expressing structure, it is sufficient to select a few keyword prefixes expressing semantics. For example, "homologs: ReceptorX" telling to search for homologs of ReceptorX. Because the term homologs is a concept of the MeSH[5] thesaurus, the system infers that in fact both, orthologs and paralogs might be important to the user.

The central issue in using structure and/or semantics to refine keyword search, is the choice of the relevant concepts. Too many make the approach inherently complicated, similar to a full-fledged structure-oriented query language like XPath or XQuery. Too little or the wrong choice of prefixes turns the approach useless. And here adaptation comes into play. Tracking queries of an individual or a group of users and hence their interests, enables the suggestion of exactly the relevant prefix refinements. Given two scientists A, B working in the same pharmacology department on toxicological models. Hence, their demographic classes are correlated. Let's assume that in the search history of B the entry "homologs:ReceptorX" exists, i.e. s/he searched once using the prefix extension *homologs*. As the system has semantic annotations, it knows that ReceptorX is an instance of the *Protein* concept. Now let's assume scientist A, having an empty search history, is using the traditional keyword search to query ReceptorZ. Again, the semantics provide a way to detect ReceptorZ as an instance of the *Protein* concept. Since the adaptive system knows that A and B are correlated, it searches for items unknown to user A. Both scientists have searched for Proteins but only B has refined the query with the prefix semantics. Thus the system suggests A to search for "homologs:ReceptorZ". In conclusion,

[5] Medical Subject Headings: http://www.nlm.nih.gov/mesh/

a recommender system could dynamically select and suggest structure and/or semantic keyword prefixes to a scientist.

3 Related work

Data integration techniques have always been of great interest for industry and research. Interesting integration approaches in the biological domain are e.g. Kleisli [Davidson 97], Tambis [Baker 98] and BioMediator [Donelson 04].

Kleisli is based on the Collection Programming Language (CPL). CPL uses sets, bags, lists and records to describe any data. In addition, it offers functions for manipulating data. Even though this approach is very powerful, it has the handicap of not offering a mediated schema over the data sources, i.e. semantics are missing. Therefore the task of choosing the appropriate data source remains at the user.

Tambis is the first semantic approach in the area of biology for data integration. It is based on the Tambis ontology (TaO) which models in parts the molecular biology domain. The TaO concepts are used to model the database sources as well as to express source independent declarative queries. Tambis uses Kleisli's CPL for accessing the data sources. Drawbacks are the limited number of databases which can be queried and the static TaO. The fundamental Tambis ontology can't be customized, making it difficult for users with different schemata to use the same system [Donelson 04].

The BioMediator approach uses an annotated mediated schema to model data sources and their relationships. A source knowledge base contains the mediated schema (which describes entities, attributes and relationships of interest to a particular group of researchers), a list of all data sources and the mapping rules. In contrast to Tambis, the mediated schema can be customized by editing with the Protégé[6] Ontology Editor. Schema adaptation requires modeling knowledge thus remaining an expert task.

The described approaches differ in degree of user guidance. While Kleisli requires the user's knowledge to decide which database to choose for querying, BioMediator makes the choice itself, i.e. the system knows more about itself and its data reservoirs. Therefore the user is unburdened in decision making. We propose to further reduce the discrepancy between the knowledge a system has about itself and a user needs to know about a system. This could be achieved by the combination of resource integration, semantic technologies, adaptive systems and an advanced query engine. The problems depicted in the introductory scenario might be solved, thus improving dramatically a scientist's information gain. If the adaptive system knows the scientists information needs it can tailor

[6] http://protege.stanford.edu/

navigation pathways to their specific requests and help by suggesting relevant extensions for keyword-based query. Therefore precision & recall of search results can be improved. In contrast to systems in the public domain, our approach addresses a closed domain, namely the corporate intranet. Here, we have the advantage, that we could use a priori knowledge about a user's roles, tasks, educational background, current department, involved projects, etc.

4 Summary and Outlook

Information overload has become a severe problem in the public domain and in companies. Traditional search and integration approaches perform poor in answering a scientist's queries. New techniques such as semantic technology offer means to apply meta information to data and resources, thus enabling computers to reason on the data. A user-centric interface to the information is still missing even though semantics have been added. Adaptation can close the gap between a user's interface and the underlying data reservoirs by customizing the communication patterns.

The extension of traditional keyword-based query with structural and/or semantic prefixes offers a simple interface for building more complex queries. The proposed combination of prefix extensions with adaptation could emerge as a useful concept for improving information access. As this idea is still in its infancy, it is the task of further research to exploit its full potential.

References

[Baker 98] P. Baker et al.: "Tambis: Transparent access to multiple bioinformatics information sources"; Proc. 6th ISMB, AAAI Press, Menlo Park, 25–34, 1998.

[Baldoni 05] M. Baldoni et al.: "Personalization for the semantic web"; Proc. Summer School Reasoning Web, LNCS, 3564, 173–212, 2005.

[Burke 02] R. Burke: "Hybrid recommender systems: Survey and experiments"; User Modeling and User-Adapted Interaction, 12(4), 331–370, 2002.

[Cohen 03] S. Cohen et al.: "Xsearch: A semantic search engine for xml"; VLDB, 2003.

[Davidson 97] S. Davidson et al.: "Biokleisli: A digital library for biomedical researchers"; Int. J. on Digital Libraries, 1(1):36–53, 1997.

[Donelson 04] L. Donelson et al.: "The biomediator system as a data integration tool to answer diverse biologic queries"; Medinfo, 72, 768–772, 2004.

[Mühlbacher 08] S. Mühlbacher: University of Regensburg / Roche Diagnostics GmbH Penzberg, Dissertation, 2008, forthcoming

[Xue 03] G. Xue et al.: "Implicit link analysis for small web search"; Proc. 26th ACM SIGIR, 56-63, 2003.