Match-Making based on Semantic Nets
The XML-based Approach of BaSeWeP

Andreas Billig, Kurt Sandkuhl

Fraunhofer-Institute for Software Engineering and Systems Engineering ISST
Mollstr. 1, D-10178 Berlin, Germany
E-Mail: [Andreas.Billig | Kurt.Sandkuhl]@isst.fhg.de

Abstract: Selection of the right information with respect to the user´s needs can be considered as one key feature of the future semantic web. This paper contributes to this issue by presenting an approach for match-making in Web-Portals between needs of a seeker and offers of a supplier. Match-making within the »BaSeWeP« framework is based on semantic nets and structured documents; XML is used for representation of content schemata, constraints, mappings and semantic nets. Basic concepts of the approach, aspects of content retrieval and application scenarios are introduced. Design issues with respect to other models from information retrieval are discussed.

1 Background

Among the activities towards the next generation of Internet, research and development work in the area of the »semantic web« has to be considered as important contribution to need-oriented information supply to Internet users. One important activity in this field is the selection and retrieval of the »right« information or service with respect to the user's needs. In our research work we are especially interested in mediation between supplier and seeker, i.e. match-making between needs of the seeker and offers of the supplier. This paper presents an approach for match-making based on structured documents and semantic nets that is implemented as a part of BaSeWeP[1], a software framework for the implementation of Web-Portals. BaSeWeP intensively uses XML [Br98] for representation of the various parts of the content specification (schemata, constraints, mappings, etc.).

The BaseWeP approach [BSW00] introduced in this paper brings together experience from component-based software development, dynamic documents based on internet-technology [MB99], meta-information systems [La97] and electronic business. Our work is integrated into the research field »information logistics“ (see below) and Web-Portal implementation projects (section 3.2).

After having discussed background information on information logistics, we will introdu-

1. BaSeWeP = Basic Support for evolutionary Web-Portals

ce basic concepts and components of BaSeWeP (chapter 2), investigate the match-making approach (section 3.1), elaborate potential application scenarios (3.2) and discuss design issues with respect to other models from information retrieval (3.3).

Information Logistics

In the field of individualisation and personalisation of information products, a wide range of technologies have been developed, including dynamic hypermedia systems [SS92], meta-information systems [HS92], content conversion [Re92] and content generation based on user profiles, and many more. All these technologies provide assistance in retrieving the right content, but apparently are not the »silver bullet« against information overflow.

As a consequence of this situation, Fraunhofer ISST has established the research field »information logistics« [DL01], funded by the State of North Rhine-Westfalia. The main objective of information logistics is development of concepts, methods, components, and solutions for a need-adequate information supply for individual users.

From an information logistics viewpoint, need-adequacy is defined by considering five dimensions

- *content* has to be tailored to the needs of the user in his specific situation
- on-*time* delivery of the right content via the right communication channel
- *presentation* format has to be tailored for available communication channels and structure of content
- *location*-awareness in information-selection and -delivery
- *quality*-parameters and quality-of-services features for content, security, presentation and communication have to be tailored to the user's needs

The BaseWeP approach introduced in chapter 2 concentrates on the content dimension of information logistics. Nevertheless, object model and application architecture are closely interrelated to information logistics reference model [Sa01].

2 BaSeWeP: Basic Concepts and Systems Architecture

BaSeWeP is a software framework for the implementation of Internet-based Web-Portals and has been developed by Fraunhofer ISST based on experience of several portal implementation projects. Basic concepts and components of the systems architecture are introduced in this chapter.

2.1 Basic Concepts

Semantic Nets

Portals, as we think of them, should administrate field or theme related content and offer it to the respective community. The concepts of the field and how they relate to each other are described with the help of a *semantic net* and constitute the basis for content classification. The most important relationship amongst concepts is the so-called *subsumes* relationship. This expresses a hierarchical order and sums up the aggregation and generalization/specialization relationship. In addition, associations can be made, placing various concepts in a named relationship.

Let us take a portal for qualification and job offers in the area of multi-media (mecomp.net, see 3.2) as an example. The following extract of the semantic net lists various concepts and their relationships[1], which are used for defining content schemata:

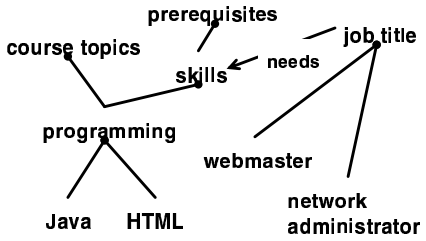

Figure 1: Extract of a semantic net from the domain »IT-qualification«

Semantic nets in BaSeWeP are not only used for stating classical relationships like „programming *is a* skill“ or „webmaster *is a* job title“ but also for stating the so-called *context* relationship , e.g. „HTML in context of programming“.

Content schemata

Amongst the portal's information suppliers are suppliers of education possibilities, on the one hand, and suppliers of jobs available, on the other hand. Respective content schemata are defined for these supplier groups. They consist of

- *concept references* for categorizing the offer in the semantic net,
- *properties* for describing the offer with the help of (structured) types and
- *resource references* for describing the offer with the help of document types[2]

Properties can possess both simple as well as structured types. To this end, the portal provides a so-called *property type pool.*

1. The edge with the filled-in circle depicts the subsumes relationship, while the arrow shows the named and specified association.
2. Documents referenced by resource references may have any kind of MIME type

Both a Java class as well as an XML-DTD fragment are available for each of these types, allowing for a representation of the content properties as a program-internal object with behavior and as an XML-stream.

```
Content Schema Job
  property location of type Address
  property start of type Date
...
  concept reference to job
  concept reference to prerequisites
...
  resource reference jobDescription
    of type postscript
```

Figure 2: Example for a content schema

Figure 2 shows an example of a content schema for job offers[1]. The resource reference of the content schema, unlike both other schema elements, defines a minimal structure. Current instances may point to resources not defined in the schema.

Figure 3 shows an instance of the content schema Job. The instance of a concept reference is a set of paths from the semantic net - limited to the subsumes relationship. The root of each concept path in the instance has to be identical with one concept reference of the schema. Instances of resource references are built using an URI (Universal Resource Identifier).

```
Content job1011 of Schema Job
  property location = [Miller Road 152, ...]
  property start = [1,1,2001]
...
  concept paths =
    { jobtitle => webmaster,
      prerequisites => skills =>
      programming => [ Html, Java ] }

  resource jobDescription =
    "http://mycompany.com/j1011.eps"
  resource companyPresentation =
    "http://mycompany.com/pres.ppt"
        of type powerpoint
```

Figure 3: Example for a content instance

1. Instead of using the XML representation for semantic nets, content schemata and instances are shown in this example using pseudo-language for the purpose of legibility.

Content-Constraint-Rules

Validation of the content serves to avoid simple inconsistencies concerning the seperate schema elements. Invalid property values or dependencies unaccounted for between these are some of the simple inconsistencies. More complex constraints allow for the editor to be informed in case of possible incorrect data or even the expulsion of content from the portal in order to ensure content quality.

The basis on which the constraints are defined is made up of operations over

- the types of the property type pools
- the set of concepts paths and
- the resource types.

To this end, the property type pool offers Java methods. Simple operations originating from the field of symbolic computing are available for formulating constraints over the concept paths, while operations on resource types come from the area of multimedia processing.

The constraint rules per content schema are defined based on these operations. Constraint rules have two parts and are controlled upon transfer of the content from the supplier to the portal. We have the condition to be satisfied, on the one hand, and the action to be carried out if the condition is not satisfied, on the other hand. The condition is a logical expression over the operations of the three schema elements.

Actions may incur the notification of the editor, perform the calculation of derived properties or calling systems services. Calling systems services may lead to new entries in the information base and to step-by-step construction of a knowledge base, which can be used for specified searching in the content as well as for content navigation. This is illustrated by the following examples for constraint rules:

```
Course Constraints
  ensure duration <= 10 and
      price.toEuro() >= 50000
  failure action
      warn ('too expensive for short course ?')

  action costLevel = price.toEuro() / duration

Job Constraints
  action
  forall p match 'prerequisites=>skills=>*' do
    addRelation needs (p, 'job title=>.')
```

Figure 4: Example for a constraint rules

The last two rules have no ensure and will be evaluated unconditionally. The third rule refers to a named association in the semantic net and ensures that all concept paths from a content instance giving the required skills for the job are set in relation to the job title. In this way, the specific content search, which goes beyond simple queries to the information base, on the one hand, and statistic evaluations or the construction of a knowledge base, on the other hand, are supported.

2.2 Systems Architecture

The systems architecture was based on requirements that derive from the development process of a web portal. The process from creating the content up to its presentation to the user is usually divided into the construction and selection phase. The selection phase covers navigation inside the content space, specific content search and content presentation. The construction phase includes creation, validation, storage and restructuring of content. For the steps creation and validation the system should offer services for a precise definition of content schemata, a declarative description of the content constraints, and an expansion to include new content schemata without redesigning the information base.

As a relational database had to be used for data storage, content mapping to a relational model had to be impelemted. Furthermore, the automatic transformation of content in the information base had to be realised in order to enable restructuring of parts of the content schema.

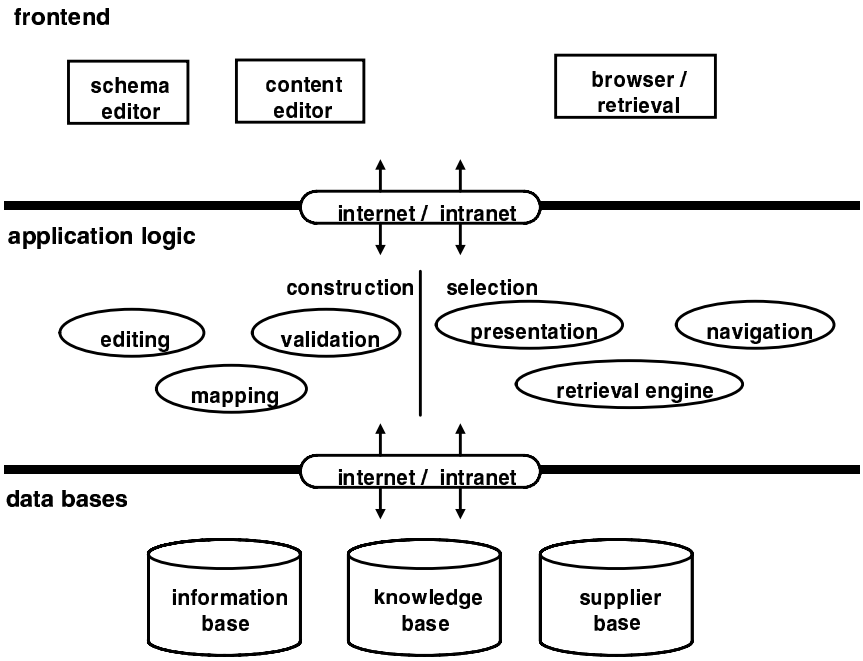

Figure 5: Systems architecture and components of BaSeWeP

In accordance with these requirements, a component-based architecture on the basis of Enterprise Java Beans has been developed. Figure 5 shows the rough architecture of a portal system which is divided into three layers: the frontend layer, containing the GUI, the application logic, which provides the services for construction and selection of the content, and the data layer.

Generation and Parameterization based on XML

The content schemata and constraint rules which are represented in XML are summarized in the so-called *Content Specification*. Additionally, the content specification includes the rules which determine the mapping of the content instances to a relational database and the presentation of the content to the user.

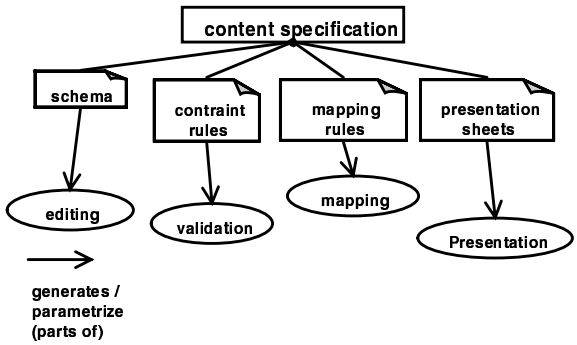

Figure 6: Use of XML specifications for parameterizing BaSeWeP

The content specification is used to generate or parametrize system components, or, more accurately, services. Due to the required flexibility, it is recommended to parametrize services during the test phase of the portal. In this case, the descriptions are read when the system is started up and kept as XML-DOMs (XML Document Object Model). Before releasing the portal, Java fragments can be generated for better performance. Figure 6 shows the connection between services and the content specification.

3 Match-Making in BaSeWeP

In the field of information retrieval, the issue of match-making has been addressed for more than a decade by numerous approaches that roughly can be classified in various types of searching algorithms and matching approaches. Searching aims at retrieving results from a given set of data items based on a search clause. Traditionally, high precision and completeness are quality parameter for search algorithms. In the context of document and content retrieval, this often is achieved by using keywords, descriptors or meta-data.

In comparison to »searching« we use the term match-making for retrieval operations where no high precision is required but the focus is on similarity of information objects. Similarity can be defined considering the structure of information, the content or the application context. Our approach focusses on similarity of the concept path sets of content schema instances.

3.1 Content-Retrieval

As with classic search engines, it is possible in BaSeWeP to search by indicating the concepts or a logical expression for these. The expressions

- (Java *and* HTML) or just
- programming

lead, e.g., to the job offer Job1011 shown in figure 3. Furthermore, it is possible to search by referring whole concept paths. This is not just to guide the user with the semantic net, but also to provide that the context information of a concept influences the search results. In order to illustrate this context based match-making, we assume an extended net where the new concept »C++« is added in the context of hardware and software programming (see figure 7).

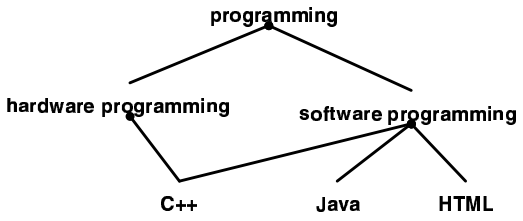

Figure 7: Example - extended sematic net

If job offers are searched for where C++ is a requirement within hardware development, the result of querying with the path

- ... => hardware programming => C++

provides a more exact result. In addition, a specific search using the conditional expressions over properties can also be carried out (figure 8).

```
retrieve c:Course
match (c, (course topics => programming =>
            software programming), s)
where Course.start < [1,11,2002]
order by s
```

Figure 8: Example query

In the preceeding example all courses which start before November and have the topics programming are retrieved. Additionally the results are ranked according to the similarity

score of the path match, i.e. courses that include the topic 'software programming' are ranked higher than courses with the topic 'hardware programming' or only 'programming' in general.

Finally the sharing of the semantic net portions between content schemata (via the category references) allows us to compare instances of different content schemata regarding similarity . For example the query

```
retrieve c:Course, j:Job, s
match (
  c[course topics => programming =>*],
  j[prerequisites => skills => programming =>*], s )
order by s
```

ranks those pairs of courses and jobs higher that have more correspondence concerning topics and skills.

3.2 Application Scenarios

The match-making approach introduced in section 3.1 has been used for validation and evaluation purposes in various application scenarios. Most scenarios were closely related to the implementation of an Internet-based Web-Portal. Web-Portals can be considered as (commercial) information products providing tailor-made information and services for communities, specific target groups or application domains.

The portal implementations using BaSeWeP´s match-making were realised in three application domains: technology transfer, education and qualification, and formation of virtual supplier organisations. In each domain mediation between supplier and seeker is based on postings: The content supplier provides structured documents with information concerning his offer. Seekers navigate in this information by using various mechanisms of the Web-Portal (browsing, keyword search, thesaurus-based functions, etc.).

In the area of *technology transfer*, the main objective of the portal was mediation between supplier and seeker of new technologies. In this scenario, the semantic net had to include terms and concepts relevant for the technologies of the domain in question, industry sectors and activity types. Content schemata were developed for innovations, products and events. This type of technology transfer portal was among others developed for the world bank (www.technologymatcher.com) and TechnologyMall Inc.

In the field of *qualification and education*, Fraunhofer ISST has implemented and currently operates the Web-Portal mecomp.net[1]. Mecomp is a portal for people interested in jobs, qualification measures and professional expertise in the area of new media, communicati-

1. http://www.mecomp.net

on and information technology. This research project is funded by the State of Berlin and is a joint activity of UdK Berlin and Fraunhofer ISST. The semantic net for this portal reflects the terminology from information and media technology that is necessary for characterising qualification measures and personal competences. Content schemata are defined for course and job offerings and personal competences.

In the context of *virtual supplier organisations*, the match-making approach currently is used and enhanced for the formation of networks of small and medium sized enterprises (SME). Current trends of globalisation and increased competition require new forms of business organisation and support. Especially in small and medium sized enterprises (SME), the competitiveness and future market position of an enterprise is closely related to the ability of cooperating with partners in SME networks or virtual supplier organisations. The formation of SME networks is based on competence models of the SMEs. The competence models are linked to product and process structures being represented as a semantic net. The approach was designed for SME networks in automotive and IT-industry and is currently evaluated in a cooperation project of European and Latin American universities [HS02].

In addition to the three application fields mentioned above, potential application scenarios have been identified in the field of health insurance (support of therapy planning by matching diagnostical information with former cases) and e-learning (generation of individualised education material based on content objects).

3.3 Discussion

The design of our match-making approach described in 3.1 is based on experience of two generations of Web-Portals. The first generation primarily was applied in the field of geological and environmental meta-information systems [La98]. Here, match-making only was based on searching meta-data and comparing them. Experience has clearly shown, that this early approach was not adequate to complex information retrieval and seeking situations. Other empirical studies, e.g. the study made by Byström and Järvelin [BJ95], confirm that the more complex the work task the less the user usually can define his or her information need. Thus, match-making approaches in complex situations as technology transfer or formation of virtual supplier organisations have to provide more sophisticated means to support seeking and retrieval.

When designing the second generation of our Web-Portal platform, we therefore took into consideration experience from information seeking strategies. Belkin et al. [Be95] introduced a four dimensional model of information seeking strategy. Part of this model is a distinction between four information seeking modes that were applied when designing BaSeWeP‘s match-making. The four modes are:

- method of searching (scanning or browsing): BaSeWeP was designed to support scanning of content objects by using the match-making mechanisms described in 3.1 and browsing step by step through the content of a portal.

- method of retrieval (specification or recognition): our match-making approach obviously is based on (unprecise) specification of the information need. Recognition only occurs when browsing through the content of a portal.
- goal of retrieval: Belkin et al. distinguish between »learning about the system« and »finding relevant information«. The application scenarios for BaSeWeP primarily require finding of information. Unprecise searching may be considered as aspect of »learning«, but this is rather a side effect than an planned design goal.
- resource considered (information object or meta-information): BaSeWeP uses information objects (= content objects) as well as meta-information for information seeking or browsing.

By observing these modes of information seeking strategy, we widened the possibilities to find information siginificantly in comparison to the first portal generation and provided enhanced support for complex information seeking scenarios.

Match making is often characterised as »searching with unprecise specification of information needs«. In this context we have to discuss the aspect of relevance of information. Ingwersen [In92] and Saracevic [Sa96] consider several types of relevance, e.g. algorithmic, topical and cognitive relevance. The underlying concepts for algorithmic relevance, i.e. the relation between the query features and the search result, and for topical relevance, i.e. releation between aboutness of content objects and query, are introduced in 3.1. Both aspects are implemented on basis of semantic nets which leads to increased flexibility for the user:

Match making in BaSeWeP is to a large extent based on comparing concept paths. From the user´s point of view, topical relevance is higher if the concept paths of his or her query are completely identical with the selected content objects than if only a subset of the nodes of the concept path are identical. The Web-Portals based on BaSeWeP therefore provide the possibility for the user to define a similarity indicator (percentage value) that specifies what degree of identity of concept paths is required for a query.

We consider the match making by comparison of concept paths and the similarity indicator as important contribution to increase cognitive relevance of the results, which is the association between perceived information need of the user and the match making results.

4 Outlook

As future work, we plan to enhance the evolution capabilities of BaSeWeP and the relevance of query results. Systems and, especially, the definitions of semantic nets, can change in course of time. As with the discipline of software engineering, it is not advisable to develop data models which don’t go beyond the "waterfall" principle and don’t include its ’evolution’.

Currently, capability for evolution in our approach is restricted to the semantic net. Modifications in the net lead to a transformation of the respective content objects. As future work, we plan to extent evolutionability to the other components of schema definition, as for example properties. In this context we currently are carrying out a feasibility study elaborating the transformation of queries. We expect to discover a number of research questions and perhaps serious problems because conserving semantics of queries while transforming them is no triviality.

In the field of relevance, we plan to add features that will support situational relevance. Situational relevance can be defined as the usefulness of objects to current interest of a user. Our opinion is that situational relevance can be increased significantly if the context of a user with respect to his or her work task, location, goal or motivation is taken into account. Recently, some research activities in the field of information logistics have been started investigating the definition of formalized models for contexts and the role of machine learning based on these formal models. These approaches could be incorporated into BaSeWeP by deveolping an interface or web services for the access on context information.

5 References

[Be95] Belkin, N.J.; Cool, C.; Stein, A.; Thiel, U.: Case, Scripts and Information Seeking Strategies: On the Design of Interactive Information Retrieval Systems. Expert Systems with Applications, Vol. 9, pp. 379-395, 1995.

[BJ95] Byström, K.; Järvelin, K.: Task Complexity Affects Information Seeking and Use. Information Processing & Management. Vol. 31, No. 2, pp. 191-214, 1995.

[Br98] Bray, T.; Paoli, J.; Sperberg-McQueen, C.M.: Extensible Markup Language (XML) 1.0, World Wide Web Consortium, Recommendation February 1998.

[BSW00] Billig, A.; Sandkuhl, K.; Wendt, A.: Basic Support for Evolutionary Web-Portals: The XML-based BaSeWeP Approach. IASTED 2000 conference proceddings, Las Vegas (USA), November 2000.

[DL01] Deiters, W; Lienemann, C.: Informationslogistik - Informationsversorgung Just-in-Time. Symposion Verlag, 2001.

[Hs92] Hsu, C.; Babin, G.; Bouziane, M.; Cheung, W.; Rattner, L.; Yee, L.; Metadatabase Modelling for Enterprise Information Integration, Journals of System Integration, Vol. 2, No. 1, February 1992, pp 5-37.

[HS02] Henoch, B.; Sandkuhl, K.: Competence Modeling as a Basis for Formation of SME-Networks - The SME-Chains Approach. Accepted paper at WWDU 2002, May 2002, Berchtesgaden, Germany; to be published in proceedings.

[In92] Ingwersen, P.: Information Retrieval Interaction. Taylor Graham, London, 1992.

[La97] Langer, T.: Definition und Nutzung multidimensionaler Datenmodelle auf relationaler Basis für flexible Metainformationssysteme. Proc. GI-Workshop »Multidimensionale Datenbanken«, Ulm, März 1997.

[La98] T. Langer: MeBro - A Framework for Metadata-Based Information Mediation. 1st Int. Workshop on Practical Information Mediation and Brokering, and the Commerce of Information on the Internet (I'MEDIAT'98), Tokyo, Japan, Sept. 1998.

[MB99] Merrall, S.; Billig, A.: Dynamo: Dynamic Document Generation in Java, The Practical Application of Java - Conference, London (UK), April 21 - April 23, 1999.

[Re92] Rearick, T.: Automating the Conversion of Text into Hypertext. In Berk, E.; Devlin, J. (Eds.): Hypertext/Hypermedia Handbook. Intertext Publications MacGraw-Hill Publishing, 1992.

[Sa01] Sandkuhl, K.: Technisches Referenzmodell für informationslogistische Anwendungen. ISST-Bericht, Fraunhofer ISST, 2001.

[Sa96] Saracevic, T.: Relevance Reconsidered ´96. In Ingwersen, P.; Pors, N. O. (eds.): Information Science: Integration in Perspective. Royal School of Library and Information Science, Copenhagen, Denmark, pp. 201-218, 1996.

[SS92] Sandkuhl, K.; Schoepf, V.: Issues and Limits of Dynamic Hypermedia Systems. Proc. Hypertext/Hypermedia ´92. Springer Verlag, Berlin, Heidelberg, 1992.