Portal Information Science: German DGI, AGI and universities build science portal

Manfred Hauer

AGI – Information Management Consultants, Germany
E-mail: manfred.hauer@agi-imc.de; URL: http://www.agi-imc.de, http://www.dgi-info.de

Abstract. Since the summer of 2004 AGI and DGI (Deutsche Gesellschaft für Informationswissenschaft und –praxis) are providing the “Portal Information Science”. It indexes tables of contents of books, articles and websites. The portal is based on AGI’s capturing and automatic indexing program intelligentCAPTURE, the semantic retrieval engine intelligentSEARCH and semantic resources in IC INDEX. The content is processed by students from the FH Burgenland in Eisenstadt (Austria). The thesaurus InfoData of FH Potsdam (Germany) and a thesaurus of students of FH Cologne are integrated in searching.

1. Introduction

Information science in Germany, Switzerland or Austria did not have a systematic electronic repository of contents published in “Information Wissenschaft und Praxis” (IWP former NFD), published in DGI conferences or other events in information science and practice. The database InfoData produced at FH Potsdam by IZ Information Science is a reference database with restricted access. This is not in coherence with current information habits of students and scientists, who want open access to full text articles with good retrieval. So has grown a gap between printed content, easily searchable content and users’ needs. This is an anachronism for information professionals building and using large online services for a long time. This is why three people at DGI and AGI decided to start with a new service, without a long discussion inside the association or between scientists and universities, setting up big business plans – getting no money at the end.

Although many college lecturers may already have a personal website, but a shared or integrated solution was missing. In 2004 there was the turning point at different places: while “Portal Information Science” was starting, open archive eLIS was invented in Spain and Italy and a little later IZ Potsdam announced a new collection of lecture material. In February 2005 representatives of each project and some universities – all universities where invited – had a meeting in Berlin for finding synergies. Now all are motivated to join forces. AGI has technology, IZ Potsdam manpower and eLIS foreign content. HI (Hochschulverband Informationswissenschaft) and DGI (Deutsche Gesellschaft für Informationswissenschaft und –praxis) deliver content in German. Springer publishing house sends bibliographic data of books.

At this point Portal Information Science presents itself. What is of interest here is not the content but the methods behind.

Since Gerard Salton it is actually clear what retrieval solutions have to look like. Users need a list of results sorted by relevance. At the latest the emergence of web search engines made that clear to
everyone. (Actually they would need validated and customized answers but technical methods are not quite that far yet.) This goal can neither be accomplished by Boolean logic nor by common, unweighted indexing. How does the “Portal Information Science” succeed?

1. Efficient acquisition of resources.
4. Collaborative production of content.

The Portal Information Science is technologically a copy of the big science portal Dandelon.com which is produced by libraries in four states today and developed by AGI. Although the same methods and technologies are used, they are applied to a rather small domain – information science.

2. Efficient acquisition of electronic resources

First intelligentCAPTURE works like a “vacuum cleaner” for capturing any relevant content and like a “continuous-flow water heater” making content hot:

- Paper: by means of scanning, OCR and PDF conversion. Content such as tables of contents and essays are prepared. E.g. one DGI conference transcript was acquired article by article via scanning. The rest of conferences was available electronically. Most tables of contents of books searchable have been scanned.

- Files: if image files are available, the processing takes place analog to paper. Otherwise a conversion to PDF by Acrobat Distiller is necessary. All common file formats can be converted to PDF. Import directories are supervised continuously and every new file is integrated into the workflows automatically. This is the way IWP and new conference volumes are captured and indexed.

- Websites: websites are summarized to one big file and processed as a single entity – just like a book or article. With two hierarchy classes within the link hierarchy there may already be 50 pages collected, with three classes it may be a few hundred. This approach is different to other search engines which index every single page separately. Within the entity a full text search is possible and every hit is announced in the Acrobat Reader at the top right frame in the portal. Yet due to copyright reasons these files often are immediately destroyed, in this case the starting URL is opened at the same place and searching is based on automatic indexing terms only. The websites spidered belong to universities, institutes, scientists, professional groups, software or service companies. The exhibitors of Online Information 2004, ComInfo 2004 and other fairs as well as some speakers’ contributions were “spidered” like this. This spidering can recur automatically in cycles, freely configurable, depending on the probable update-speed of the destination address. Further filters are to be worked on.

- For 2005 objects are to be made describable through talked language which is converted into text by speech recognition, e.g. pictures, videos, music and language or textless objects like objects of art or collectors’ items.

- Finally structured files – XML or other formats – can be imported and prepared for further processing. Soon files of Springer will be loaded – 1500 book reference data in information retrieval.

The punch-line of all these inputs is an efficient and respectively specific workflow. All input-workflows can run at the same time and be worked off simultaneously. This processing can be carried out only on one workstation or in case of a high amount of data on several computers in parallel. A server program
supervises the complete production of every workstation involved and distributes the computer loading. For this an IBM Lotus Domino server is necessary and IBM Lotus Notes clients at each computer, as all modules are embedded in the cheap and very efficient client program of IBM.

For 2005 numerous expansions are planned, a considerably more profound, automated analysis of the document structures which subsequently are included in XML. Even faster and better OCR results, more languages and also the recognition of old documents with fracture font are in work. The filtering on content-relevant text sections will be improved, e.g. to extract and process only sections which deal with certain persons, institutions, themes or products.

3. Machine indexing

Computer scientists, technically oriented information scientists, computer linguists and mathematicians have struggled for efficient indexing methods for a long time. The methods described above deliver the "raw material" for indexing. Since 1999 AGI uses the indexing solution of IAI in Saarbrücken, called AUTINDEX. It is integrated in intelligentCAPTURE and named there CAI-engine (Computer...
Fig. 2. One term (grey) is always in the center, non-descriptors (BF), narrower terms abstract (UBA), related terms (VB) and translation (TRA) are represented here. On the left, in the section “Ähnliche Begriffe/similar concepts” are automatically displayed based on root searching if any is available. Above “Ihre neuen Suchbegriffe/Your new search terms” can be collected. Each term can be added to the search list one by one or as a group by double click the red ball. Each term can have more details in the message box like scope notes, definitions or URLs. At top left section back- and forward rotation and moving is available. Each green ball is a starting point for next visualization.

Aided Indexing). The engine combines several methods: morphological ones for recognition of the basic forms, syntactic ones for recognition of word families, a rough thematic classification of nouns, heuristic ones for the text structure analysis, statistical ones for term weighting, then a comparison of the words and phrases classified as relevant by thesaurus or classification, there is a recognition of names of persons, organizations and geographical regions and finally even an optional translation of descriptors according to thesaurus implemented. The complete recognition is language specific and must be prepared for every natural language with morphological dictionaries, sets of rules/grammars and domain specific thesauri. In practical use there are German and English so far. Other languages are in work more or less extensively. A large more simple but domain-independent IAI thesauri is available for German (250,000 terms) and for English (500,000 terms).

A modern Windows computer needs 20 to 40 minutes to process a whole DGI conference (350 pages/about 40 articles) from input in intelligentCAPTURE as files up to output to intelligentSEARCH databases or other catalogs. The manual typing of bibliographic data takes longer.
At the DGI conference in 2003 the tool was explained in detail – of course the full text is available on portal – search keyword “AUTINDEX” in Artikel/articles section [1].

Within intelligentCAPUTURE the process of indexing is fully automated. Although it is possible to carry out changes at the result manually. But it is far more efficient to improve the semantic base instead, the “rake” the text is combed through with. At the same time the CAI-engine itself is developed continuously and a new, English version is available now.

A statistical tool IC CATE finds terms suitable for a thesaurus update in an indexed document collection. This is an application of the Zipf’sches law, but related to selected, linguistically standardized free descriptors and term weighting of CAI-engine. The decision to transfer a terms into the thesaurus and to network correctly always rests on the thesaurus editors. Many terms were selected in this way for updating of InfoData thesaurus edited in IC INDEX. The updated thesaurus later can be export and compiled for CAI-engine. Parallel current version of the thesaurus is visible any time online – use in the portal button “Topic Maps Visualization” for browsing.

IC Index is a classic among the thesaurus development programs and furthermore supports the multilingualism by integrated machine translation and powerful translation workflows which is always finally checked by human. It supports per default 26 different relationships and suits well to globally distributed editor teams.

4. Semantic retrieval with ranking

The interface to the end-user is designed similar to Google consciously since a familiarity can be presupposed here.

The Lotus Domino server has built in IBM’s GTR (Global Text Retrieval) engine. It is an n-Gram-engine which is suitable for all languages and able to sort documents by relevance. Field search, ranges of values, Boolean operators, context operators, fuzzy search, exact spelling, enquiries with ranking values and of course parallel left, middle and right hand truncation are offered. A part of these options is documented in the online Hilfe (help) in the portal. The rest of features are used in an invisible way.

A full text index can be built up for one or multiple databases. The frequency and position in text of the search term is the basis of the relevance ranking. When passing over data from intelligentCAPUTURE to intelligentSEARCH the indexing termweights of CAI-Engine is used to steer ranking of GTR.

The author of this article tested with 99 students how different students describe the same document. Groups with three students each subscripted the same text and afterwards compared the degree of accordance. The average value is 21% – means 4 of 5 indexing terms are different. If you transmit this to creating queries, then there can be a match between queries and documents of rather less than 20% only if queries are not made “more intelligent”. Librarians and documentation specialists counter this with classifications and thesauri by delimiting the permitted language space. However, considering the query possibilities on the internet, where few people work with controlled indexing languages, especially such terminological systems can be used for expanding queries. Therefore the portal dandelon.com, based on intelligentSEARCH has implemented currently 415,000 topics, networked terms, which extend the user’s search term by synonyms, translations and narrower terms. Portal Information Science is using about 10,000 topics in three thesauri. The user can optionally include all other relation types in the wide field of association by the topic map-navigation.

However, a sorting of search results by ranking only is not optimal for every inquiry. Therefore the search result can alternatively be sorted by author or year. A topic sorting is being revised at present.
Doing retrieval tests in comparison with library systems (with Boolean logic and classical indexation) the concept of intelligentSEARCH performed considerably better [2].

5. Collaborative production of contents

AGI – Information Management Consultants, Vorarlberger Landesbibliothek in Austria and a project team of female students of the FH Burgenland inserted content. It is intended to involve further libraries, universities or documentation centers. The new coalition established in Berlin may bring in additional resources and a improved multilingual thesaurus. The portal wants to be an international one. German, English and soon French and Italian speaking content can be processed. Everyone who does not have intelligentCAPTURE software on his own can send links, CDs, attachments or printed media to portal editors, the editors are increasing the scope step by step.

Unfortunately there is not yet any funding or income related to the portal, so all work is honorary.

89. Deutscher Bibliothekertag in Freiburg im Breisgau 1999 : warenlos in die Zukunft / Buch- und Bibliothekswesen, Kongressberichte, Deutschland, Bibliothekswesen

Information und Öffentlichkeit : 1. Gemeinsamer Kongress der Bundesverbind der Deutschen Bibliotheksverbände ... zudem 39. Bibliothekertag ... Buch- und Bibliothekswesen, Kongressberichte, Information und Dokumentation, Kongressberichte

Fig. 4. The phrase “dokumentationssprachen bibliotheken” was searched here, but is not a existing phrase and so there is no hit. Next search terms automatically analyzed, both are in plural – which is not conform with German thesaurus rules. The language analyzer converts the term to singular and finds matches in thesauri. As well an AND was put between both topics – both have the same weight and must occur in the documents found. But with search strategy “Medium” recall or medium precision automatically synonyms, translations and narrower terms are added and put to the final query.

Search (user input)
dokumentationssprachen bibliotheken

Search (calculated)
(("dokumentationssprachen") Or ("dokumentationssprache" OR "Hierarchie" OR "Thesaurus" OR "Anordnungstechnik" OR "Klassifikation" OR "Ontologie" OR "Taxonomie" OR "Topic Maps" OR "Ordnungssystem" OR "Kontrolliertes Vokabular" OR "Indexing Languages") And ("bibliotheken") Or ("bibliothek" OR "Library" OR "Mediothek" OR "Landesbibliothek" OR "Hochschulbibliothek" OR "Depotbibliothek" OR "Spezialbibliothek" OR "Archiv" OR "Körperschaft" OR "Nationalbibliothek" OR "Öffentliche Bibliothek"))

Project team at FH Burgenland in Eisenstadt:
Katrin Ferschitz (Katrin.Ferschitz@fh-burgenland.at), Daniela Schmiedl, Kriemhild Leitner, Anja Prassl, Christa Scharaditsch

Thesaurus from FH Colgne: Nachrelationierung of the SWD
Daniela Trunk (danielatrunk@gmx.de) and Sarah Hartmann
Fig. 5. Document display: in left frame bibliographic and indexing with highlighting is displayed. On the right frame a PDF or webpage is loading automatically. Actions for switching to library catalog, ordering the item at a bookseller, "more like this" button and a forwarding button are available.

References
