**In:** Eibl, Maximilian; Wolff, Christian; Womser-Hacker, Christa (Hg.): Designing Information Systems. Festschrift für Jürgen Krause. Konstanz: UVK Verlagsgesellschaft mbH, 2005. S. 59 – 70.

# Information Work in the Internet Age: Localizing, Evaluating and Representing Resources

*Thomas Mandl*

**Abstract**

The amount of data on the internet continues to grow. At the same time, funding for professional information work for collection building, quality assessment and content representation is reduced. The Layer Model responds to these challenges associated with the internet. The Layer Model shows how information providers can restrict the information work dedicated to content representation to small collections and, at the same time, allows users to access much larger collections. Transfer modules apply intellectual information work to documents outside a core. This article shows how the Layer Model can be extended in order to integrate automatic quality assessment and automated collection building.

## 1 Information Services in the Internet Age

The explosion of online information and the technological advance offer great opportunities for the information industry. At the same time, however, they paradoxically pose a threat to many business models. The following opinions often expressed by users sum up some of these threats:

- Everything is available for free. Why should anyone pay for content, access or intellectual information work?
- If something is online, then it must be accessible. Intellectual information work seems unnecessary.
- I can find everything by myself. The availability of free information tools and their ease of use makes less experienced users believe that information retrieval and access are solved problems.

The availability of free information tools which are based on a large amount of data leads to satisfying results in many cases. One example are internet search engines. However, information professionals have pointed out the low overall retrieval quality of general purpose search engines and the lack of advanced search options they provide. The retrieval quality of these tools seems to be far behind the quality of systems optimized for a specific task, as comparisons have shown (Hawking 2000). The continuing growth of the internet makes information retrieval and access ever more troublesome. The task of information centers or information units in enterprises is becoming more difficult. In addition, the opinions stated above create growing pressure because intellectual information work seems obsolete and expensive.

The Layer Model proposed by Krause 1998 responds to these challenges (see also Krause 2003). It is a sketch of a new role model for scientific information centers supported by public funds. Within the Layer Model, special attention is given to intellectual content representation. The Layer Model shifts the role of information centers from a monolithic center to a flexible moderator which integrates other information collections into its service. Although these collections may not be indexed by the same staff and according to the same guidelines, they may provide an advantage for the user. As stated by Krause, there will always be information providers who will not concord to the standards proposed by central information centers, however, users will not accept that this data is neglected (Krause 2003:5). In the Layer Model, a moderator organizes layers of decreasing indexing representation quality and allows transparent access to all documents.

The Layer Model may be applied by a large variety of information providers. Intellectual indexing seems to be an outdated strategy, however, in the internet age and even faced with the ubiquitous availability of automatic indexing, it seems to flourish as never before. Information work is done by many services of the new economy, often without being recognized as such. So far, the implementation of the Layer Model has been mainly based on text categorization, a well established research area within information retrieval which assigns documents to predefined categories. These categories for an ontology in a domain are organized according to content. Text categorization uses evidence from the content of the information object for the mapping to these categories.

However, information centers carry out much more work than indexing. Especially important are collection building and quality control. In these areas, similar problems as in the case of indexing are apparent. The information centers (and many information companies) are experts for the tasks of collecting, evaluating and organizing information. In the internet age, the demand for these tasks is exploding. Paradoxically, the information specialists will not carry out these tasks any longer or at least not in the same form or to the same extent due to economic and social circumstances.

High availability of information and information technology creates a huge demand for intellectual information work, however, it also feeds the myth that this work is not necessary and it also supports the truth that this amount of work simply cannot be done. Information work is not necessary because everything is available and it cannot be accomplished because too much information is being published. The creation and the maintenance of collections as well as quality control should therefore be also subjected to the Layer Model. Obviously, there exists a correlation between the quality of an object and the quality of its content representation. However, not all information providers worry about content representation. Many small publishers of high-quality documents are not even aware of the problems of proper content representation. They sometimes assume that general purpose search engines will be the predominant referrer to their information. And even when the content is organized, the structure may be in an inaccessible format.

The next section will review the Layer Model and first implementation steps. The following two sections will deal with collection building and quality control respectively. They will both review previous work and discuss it within the framework of the Layer Model.

## 2 The Layer Model for Information Services

The Layer Model (Krause 1998) is concerned with the content representation of documents. Contrary to the traditional model of information provision, it does not assume that all information objects for one scientific discipline are indexed by the same staff and according to the same guidelines of one center. Such approaches to create uniform collections and representations have become to inflexible to cope with the reality of information provision in the internet age. The technological advance allows many other players to provide access to specialized collections and to create representations based on new representation schemes or automatic indexing. The Layer Model postulates that an information service center needs to provide access to such collections in order to optimally serve its users (Krause 1998).

The Layer Model aims at a new mix of manual indexing and automatic indexing. Because knowledge work in an information center can only be applied to a small subset of documents, this knowledge work needs to be automatically transferred to the other documents. Even when documents are indexed manually, they may not be represented within the same framework. The indexer of a different information provider may have used another ontology grounded in a different context.

The Layer Model proposes layers of documents with similar levels of content representation. Potentially, any internet document can be included and a full text index may serve as its content representation. The user can access all layers or only layers with a guaranteed level of content representation. In that manner, the Layer Model manages to integrate a large number of documents from the web without neglecting and losing high quality indexing done at information centers (Krause 1998). A similar situation occurs in many intranets where different levels of content representation can be easily identified. In a Layer Model approach, internet documents can be added as an additional layer of lower representation quality.

## 2.1 Semantic Heterogeneity

The use of different ontologies results in semantic heterogeneity (Mandl & Womser-Hacker 2001). Even identical indexing terms may have a completely different meaning because they occur in a different context in two ontologies. Some of the challenges can be summarized as follows:

- Different terminology
- Different levels of detail or abstraction
- Different order of hierarchy construction (for example: *physics -> applications of physics –> applications of radiology* vs. physics -> *radiology –> applications of radiology*)
- Different concepts resulting in different clusters

Semantic heterogeneity has been recognized as a challenge in many information retrieval applications (Chen 1998). Recently, it has also led to novel user interfaces for browsing (Heinz et al. 2003). Common approaches for treating heterogeneity are briefly reviewed in the following subsection.

## 2.2 Text Categorization

In order to allow access to an object based on an ontology, a system needs a representation of the object within this ontology. By extending the reach of ontologies beyond their primary objects, the Layer Model needs to generate representations of an object within these ontologies. This could be done by intellectual indexing, however, that would require too many resources. Consequently, automatic or semi-automatic methods need to be employed.

The intellectual work can be applied to the ontologies instead of the objects. Indexers could define mappings between the concepts or terms of heterogeneous ontologies (Hellweg et al. 2001). The mapping would define for example that term A from ontology Z is equivalent to term B from ontology Y. The generation of such concordances requires some resources, proves difficult for partial overlaps and cannot account for all interdependencies between concepts.

The most appropriate technology to delegate this task to machines is text categorization between ontologies or terminologies. In most cases, text categorization assigns documents to predefined categories based on a full text analysis (Mandl 2001, Joachims 2002). The text is indexed by standard information retrieval methods and represented by weights assigned to words or terms based on their frequency of occurrence. These terms can be regarded as features. In the same manner, terms from a controlled vocabulary like an ontology can serve as features for a text. Thus, the task for text categorization based on full text terms is equivalent to text categorization based on descriptors from ontologies. Different learning methods have been applied to text categorization. Most often, statistical association measures like Naive Bayes provide mappings between pairs of terms. These learning algorithms derive the knowledge from examples provided as training data and do not rely on further human contributions. Neural networks and support vector machines have been employed as well (Mandl 2001, Joachims 2002).

Text categorization is often used for the treatment of semantic heterogeneity. It requires that some objects have been indexed with two ontologies in order to derive a mapping. When such a corpus is not available, heuristic approaches are necessary. A two-step method was introduced by Mandl 1999. At first, a full text index of the documents is generated. Then, the terms of one ontology are searched within the documents and their occurrence is considered as evidence for a relationship between the ontology term and the terms from another ontology manually assigned to the document.

Recent technological development within the area of the semantic web offers new opportunities for the Layer Model. The semantic web attempts to set standards for the semantic markup of information in order to make online information understandable for machines. In the semantic web, many information sources are available when processing in database-like queries and for complex reasoning processes. The standards for achieving this goal stress the importance of ontologies for information management processes (Fensel et al. 2003). The vision of a full functional semantic web may be far ahead. However, the standards seem to fit the needs of the Layer Model as well. The treatment of semantic heterogeneity has already been implemented with semantic web technology in a few cases (e.g. Doan et al. 2002, Kölle et al. 2004).

## 3 Quality Models

The heterogeneous quality of documents on the internet has been a matter of growing concern. Although there is no consensus on what quality means, there is wide consensus that the quality of internet documents varies greatly. At first, the lack of control concerning content as well as form and presentation within the internet has led to the enormous success of the web. Everybody has the possibility to publish any information. This situation has also led to many documents of questionable or dubious quality among the documents online[1].

As a reaction, the library and information science field created many criteria lists for evaluating the quality of resources[2]. These lists provide indicators for the quality of web resources. However, these criteria are very difficult to apply for the average user and they require considerable resources. The criteria lists show that even intellectual quality assessment is extremely difficult. Even more so, quality control has become impossible due to the large amount of documents online. As a consequence, quality assessment needs to be partially delegated to information systems. This trend has been stimulated especially by the search engine Google[3], which has integrated quality assessment into its result rankings. Meanwhile, other systems followed the trend to incorporate quality analysis.

The following section reviews link-analysis and its shortcomings. The next section briefly shows advanced quality models and their modules and discusses their integration into the Layer Model.

### 3.1 Link-based Authority Measures

The most widely adopted approach to heuristically measure the quality of a page has been link analysis. The number of links pointing to a page are considered as the main quality indicator. A large number of algorithms for link analysis have been developed. The most well known ones are probably the PageRank algorithm and its variants (Dhyani et al. 2002).

[1] „The simplicity of creating and publishing web pages results in a large fraction of low quality web pages“ (Page et al. 1998:2)

[2] Wilkinson, G. L., Bennett, L. und Oliver, K. (1997). ”Evaluation criteria and indicators of quality of Internet resources.” In: Educational Technology, 37(3).
Librarians’ Index to the Internet (ed.) (2004). Lii.org Selection Criteria. http://lii.org/search/file/pubcriteria [Zugriff September 2004].
Fenton, S. (1997). Information Quality: Is the truth out there? University of North Carolina Chapel Hill. http://ils.unc.edu/~fents/310/ [Zugriff September 2004].
California State University Stanislaus. University Library (ed.). (2003). Evaluation of Web Resources. http://www.library.csustan.edu/lboyer/webeval/webeval.htm [Zugriff September 2004].

[3] Google Inc. (2004). Google Search Engine Homepage. http://www.google.com [Zugriff September 2004].

The basic assumption of PageRank and similar approaches is that the number of in- or back-links of a web page can be used as a measure for the popularity and consequently for the authority of a page (Page et al. 1998). PageRank assigns an authority value to each web page which is primarily a function of its in-links. Additionally, it assumes that links from pages with high authority should be weighed higher and should result in higher authority for the receiving page. The algorithm is carried out in several iterations until the result converges. PageRank can also be interpreted as an iterative matrix operation which results in an approximation of the eigenvector of the connectivity matrix of the web pages considered. Similar measures have been used for decades in bibliometrics for the evaluation of scientific literature within the network of citations (Choo et al. 2000).

Link analysis has several serious shortcomings. Certainly, quality is not the only reason for setting a link. The assignment of links is a social process leading to remarkable global patterns. The number of in-links for a web page follows a power law distribution (Dill et al. 2001). That means, many pages have few in-links while few pages have an extremely high number of in-links. In such a distribution, the median value is much lower than the average. This finding indicates that web page authors choose the web sites they link to without a thorough quality evaluation. Much rather, they act according to economic principles and invest as little time as possible for their selection. As a consequence, social actors in networks rely on the preferences of other actors (Pennock et al. 2002). A structure bias for the number of in-links of a page is also evident. Pages lower in a hierarchy are much less likely to receive in-links than homepages (Mandl 2003).

## 3.2 Advanced Quality Measures

Link-analysis considers only one knowledge source and cannot be a reliable quality measure. As a consequence, several research prototypes are experimenting with advanced quality models. One approach to measure the quality of the usability of web sites is to check the presence of tags and compare the tag structure to guidelines. For example, such programs check whether an alternative text for a picture is provided (Brajnik 2001). These systems go little beyond a HTML syntax checking and measure solely one aspect of quality.

An approach to measure the quality of text comes from educational information systems and grades the essays of students (Foltz et al. 1999). The quality of an essay is either measured by its similarity to a model essay or by its internal coherence calculated as the similarity between subsections of the essay. The distance is calculated by latent semantic indexing (Foltz et al. 1999).

Current research is intensively working on the implementation of advanced quality models which consider more parameters than links and which include several perspectives. Examples are WebTango (Ivory & Hearst 2002), Bloodhound (Chi et al. 2003) and AQUAINT[4] (Automatic Quality Assessment for Internet Resources, Mandl 2002).

The quality of documents independent from representation quality is an important dimension for the users of retrieval systems. Intellectual quality control and assessment cannot be provided for all documents on the web. Link analysis is a simple heuristic for quality assessment which has various limitations. Advanced quality assessment strategies are based on human quality judgments. They identify patterns within large pools of human quality judgments and apply these patterns to documents which have not been judged. However, the identification of reliable quality indicators remains an unsolved research question.

Overall, the situation is similar to the scenario of the Layer Model where information work cannot reach all documents. As a consequence, human information work is exploited beyond its initial purpose and transferred to a larger corpus of documents. Quality control is delegated to machines, however, the training data is provided by humans. The Layer Model should integrate content quality as an additional dimension. The quality of documents can be evaluated and they can be assigned to layers of heterogeneous quality. The user could then be offered a parameter to control the quality of the results.

### 3.3 Collection Building

To assure high quality of a collection, information centers incorporate only hand-picked resources or they rely on the quality assurance of other institutions like editorial boards or publishers. Such an approach is not well suited for the internet and as a consequence, general purpose search engines on the web take a radically different approach. They collect all pages available. The internet continues to grow very rapidly. Large search engines claim to have indexed more than two billion pages. While size continues to be a major criteria for the evaluation of search engines, there seems to be a growing trend toward the other direction. Some search engines no longer try to index as many resources as possible. Rather, they focus on quality resources or on resources for specific topics.

[4] AQUAINT is supported by the Deutsche Forschungsgemeinschaft (DFG, grant nr. MA 2411/3-1)

The first step toward the goal of automatically building high quality collections is the elimination of spam. Pages containing misleading information about their content for indexing purposes are usually considered to be of low quality. The next step may be the direction of a crawler[5] toward high quality resources. The quality of pages can be assessed by link-analysis and pages with higher quality are visited earlier in the crawl (Menczer et al. 2001). In addition, crawling can be focused on thematically similar pages. For that purpose, link and content information are combined. Pages downloaded by a crawler are analyzed, and the similarity of their content to the desired content profile is calculated. Content analysis methods known from information retrieval are applied to the similarity calculation.

A collection of topically related pages can also be interpreted as a community. The recognition of web communities has also drawn considerable interest. Communities are often identified by the link patterns between their pages (Gibson et al. 1998). An overview of collection building (also referred to as topic distillation or resource discovery) is provided in Chakrabarti 1999.

The application of collection building techniques within the context of the Layer Model can be used to create additional layers fully automatically. For higher layers, the quality requirements can be strict in order to create high quality collections. For lower layers, they can be relaxed in order to include more documents on lower levels.

## 4 Conclusion

This article discussed some of the information access challenges of the internet age. The growing amount of information creates a great need for intellectual information work. At the same time, funding for intellectual information work is not increased. As a consequence, a smaller fraction of all documents is subjected to information work. New models should better exploit the results of human knowledge work. The Layer Model is an approach for flexibly increasing the amount of information accessible and, at the same time, decreasing the quality of the content representation. Apart from automatic indexing, the information explosion has also led to approaches for the automatic construction of collections based on different policies and methods to automatically assess the quality of documents. Such modules need to be incorporated into the Layer Model. Systems should be enabled to automatically transfer collection policies, quality evaluation criteria, topical organization as well as

[5] A crawler is a program which automatically downloads internet pages and passes them on to the indexing module of a search engine.

content representation to new resources. Two objectives are reached by such an approach:

- Knowledge work is exploited beyond its original purpose and transferred to other documents
- Users gain flexible access to larger collections which are segmented into distinct layers

The necessary competence profile of future information workers is shifting. The core of typical curricula for information professionals focused on knowledge in the tradition of library science. In the near future, technological knowledge to apply the novel techniques discussed in this paper will be necessary. In addition, information management approaches in information centers and in information businesses will need to be modified to take the new division of information work between human and machine into account. Some curricula have already been adopted toward such a profile (e.g. Beneke et al. 1999).

## 5 References

Beneke, J.; Hauenschild, C.; Womser-Hacker, C. (1999). "Der Studiengang Internationales Informationsmanagement an der Universität Hildesheim." In: Ockenfeld, M.; Mantwill, G. (eds.): Information und Region. 51. Jahrestagung der Deutschen Gesellschaft für Informationswissenschaft und Informationspraxis e.V. (DGI), 181-91.

Brajnik, G. (2001). “Towards valid quality models for websites.” In: Proceedings of the 7th Conference on Human Factors & the Web (HFWEB). http://www.dimi.uniud.it/~giorgio/ papers/hfweb01.html [Zugriff September 2004].

Chakrabarti, S. (1999). “Recent Results in Automatic Web Resource Discovery.” In: ACM Computing Surveys 31 (4).

Chen, H.; Martinez, J.; Kirchhoff, A.; Ng, T.; Schatz, B. (1998). “Alleviating Search Uncertainty through Concept Associations: Automatic Indexing, Co-Occurrence Analysis, and Parallel Computing.” In: Journal of the American Society for Information Science JASIS 49(3), 206-16.

Chi, E.; Rosien, A.; Supattanasiri, G.; et al. (2003). “The Bloodhound Project: Usability Issues Using the InfoScentTM Simulator.” In: Proceedings ACM Conference on Human Factors in Computing Systems, 505-12.

Choo, C.; Detlor, B.; Turnbull, D. (2000). Web Work: Information Seeking and Knowledge Work on the World Wide Web. Dordrecht et al.: Kluwer.

Dhyani, D.; Ng, W.; Bhowmick, S. (2002). “A Survey of Web Metrics.” In: ACM Coumputing Surveys 34 (4), 469-503.

Dill, S.; Kumar, R.; McCurley, et al. (2001). “Self-Similarity in the web.” In: Proceedings 27th International Conference on Very Large Databases (VLDB).

Doan, A.; Madhavan, J.; Domingos, P.; Halevy, A. (2002). “Learning to Map between Ontologies on the Semantic Web.” In: Proceedings of the World Wide Web

Conference. http://www2002.org/CDROM/refereed/232/index.html [Zugriff September 2004].

Fensel, D.; Hendler, J.; Liebermann, H.; Wahlster, W. (2003) (eds.). Spinning the Semantic Web. Bringing the World Wide Web to Its Full Potential. Cambridge/London: MIT Press.

Foltz, P.W.; Laham, D.; Landauer, T. (1999). “The Intelligent Essay Assessor: Applications to Educational Technology.” In: Interactive Multimedia Electronic Journal of Computer-Enhanced Learning. http://imej.wfu.edu/articles/1999/2/04/printver.asp [Zugriff September 2004].

Gibson, D.; Kleinberg, J.; Raghavan, P. (1998). “Inferring Web Communities from Link Topology.” In: Proceedings 9th ACM Conference on Hypertext and Hypermedia. http://citeseer.nj.nec.com/gibson98inferring.html [Zugriff September 2004].

Hawking. D. (2000). “Overview of the TREC-9 Web Track.” In: Voorhees, E.; Harman, D. (eds.): The Ninth Text REtrieval Conference (TREC 9). NIST Special Publication 500-249. http://trec.nist.gov/pubs/trec9/t9_proceedings.html [Zugriff September 2004].

Heinz, S.; Mandl, T.; Womser-Hacker, C. (2003). “Implementation and Evaluation of a Virtual Library Shelf for Information Science Content.” In: Digital Libraries, Advanced Methods and Technologies, Digital Collections. Proceedings of the fifth National Russian Research Conference (RCDL). St. Petersburg. 20.-31. Okt. 2003. Saint Petersburg State University Press. S. 117-123.

Hellweg, H.; Krause, J.; Mandl, T.; et al. (2001). Treatment of Semantic Heterogeneity in Information Retrieval. Technical report Nr. 23, IZ Sozialwissenschaften, Bonn, Germany. http://www.gesis.org/Publikationen/Berichte/IZ_Arbeitsberichte/index.htm#ab23 [Zugriff September 2004].

Ivory, M.; Hearst, M. (2002). “Statistical Profiles of Highly-Rated Sites.” In: Proceedings ACM CHI 2002. Conference on Human Factors in Computing Systems. New York: ACM Press.

Joachims, T. (2002). Learning Text Classifiers wit Support Vector Machines. Dordrecht: Kluwer.

Kölle, Ralph; Mandl, Thomas; Schneider, René; Strötgen, Robert (2004): Weiterentwicklung des virtuellen Bibliotheksregal MyShelf mit semantic web Technologie: Erste Erfahrungen mit informationswissenschaftlichen Inhalten. In: Ockenfeld, Marlies (Hrsg.): Information Professional 2011: Strategien – Allianzen – Netzwerke. Proceedings 26. DGI Online-Tagung. Frankfurt a.M. 15.-17. Juni. S. 111-124.

Krause, J. (1998). “Innovative Current Research Information Systems in the Information Society.” In: CRIS ´98 Current Research Information Systems. Luxemburg, ftp://ftp.cordis.lu/pub/cybercafe/docs/krause.zip [Zugriff September 2004].

Krause, J. (2003). "Standardisierung von der Heterogenität her denken – Zum Entwicklungsstand bilateraler Transferkomponenten für digitale Fachbibliotheken." Technical report Nr. 23, IZ Sozialwissenschaften, Bonn, Germany. http://www.gesis.org/Publikationen/Berichte/IZ_Arbeitsberichte/index.htm#ab28 [Zugriff September 2004].

Mandl, T. (1999). "Effiziente Implementierung von statistischen Assoziationen im Text-Retrieval." In: Ockenfeld, M.; Mantwill, G. (eds.): Information und Region. 51.

Jahrestagung der Deutschen Gesellschaft für Informationswissenschaft und Informationspraxis e.V. (DGI), 159-72.

Mandl, T. (2002). Projekt AQUAINT. http://www.uni-hildesheim.de/~mandl/Forschung/Aquaint/index.html [Zugriff September 2004].

Mandl, T. (2003). “Link Analysis and Site Structure in Information Retrieval.” In: Dittrich, K.; König, W.; Oberweis, A.; et al. (eds.) (2003). Informatik 2003: Beiträge der 33. Jahrestagung der Gesellschaft für Informatik. [LNI P-35], 262-67.

Mandl, T.; Womser-Hacker, C. (2001). "Fusion Approaches for Mappings Between Heterogeneous Ontologies." In: Research and Advanced Technology for Digital Libraries: 5th European Conference (ECDL) Darmstadt Sept. 4.-8. Berlin et al.: Springer [LNCS 2163]. 83-94.

Menczer, F.; Pant, G.; Srinivasan, P.; Ruiz, M. (2001). “Evaluating Topic-Driven Web Crawlers.” In: Proceedings of the ACM SIGIR Conference. 241-49.

Page L.; Brin. S.; Motwani. R.; Winograd. T. (1998). The PageRank Citation Ranking: Bringing Order to the Web. Manuscript. http://citeseer.nj.nec.com/page98pagerank.html [Zugriff September 2004].

Pennock, D.; Flake, G.; Lawrence, S.; et al. (2002). “Winners don’t take all: Characterizing the competition for links on the web.” In: Proceedings National Academy of Sciences 99 (8), 5207–11. http://modelingtheweb.com/modelingtheweb.pdf [Zugriff September 2004].