Sérgio Tenreiro de Magalhães
Leonel Santos
Maximilian Stempfhuber
Liv Fugl
Bo Alrø
(Eds.)

CRIS-IR 2006

Proceedings of the International Workshop on Information Retrieval on Current Research Information Systems

Copenhagen, Denmark, 9th November 2006

Sérgio Tenreiro de Magalhães
Leonel Santos
Maximilian Stempfhuber
Liv Fugl
Bo Alrø
(Eds.)

CRIS-IR 2006

Proceedings of the International Workshop on Information Retrieval on Current Research Information Systems

Copenhagen, Denmark, 9th November 2006

Editors

Sérgio Tenreiro de Magalhães
Department of information Systems
University of Minho,
Campus de Azurém
4800-058 Guimarães, Portugal

Leonel Santos
Department of information Systems
University of Minho,
Campus de Azurém
4800-058 Guimarães, Portugal

Maximilian Stempfhuber
GESIS / Social Science Information Centre (IZ)
Lennéstr. 30,
53113 Bonn, Germany

Liv Fugl
Produktionstorvet,
building 425
DK-2800 Kgs. Lyngby, Denmark

Bo Alrø
Atira A/S
Niels Jernes Vej 10
9220 Aalborg Oest , Denmark

Printed by CHAPA5

ISBN: 978-972-98921-6-5

Organizing Committee

Maximilian Stempfhuber	(GESIS/Social Science Information Institute, Deutschland)
S. Tenreiro de Magalhães	(University of Minho, Portugal)
Liv Fugl	(Center for Knowledge Technology, Denmark)
Bo Alroe	(Atira, Denmark)

Programme Committee (PC)

PC Co-chairs:

Leonel Santos	(University of Minho)
Maximilian Stempfhuber	(GESIS/Social Science Information Institute)

PC members

Leonel Santos	(University of Minho, Portugal)
Maximilian Stempfhuber	(GESIS/Social Science Information Institute, Deutschland)
Dominik Sleyak	(University of Regina, Canada)
Do-Wan Kim	(Paichai University, South Korea)
Kenneth Revett	(University of Westminster, U.K.)
Liv Fugl	(Center for Knowledge Technology, Denmark)
Luis Amaral	(University of Minho, Portugal)
Manuel Filipe Santos	(University of Minho, Portugal)
Maria Flamínia Ramos	(Fundação para a Ciência e Tecnologia, Portugal)
Roberto Pacheco	(Stela Inst./ Federal University of Sta. Catarina, Brasil)
S. Tenreiro de Magalhães	(University of Minho, Portugal)
Thomas Mandl	(University of Hildesheim, Deutschland)
Vivien Petras	(University of California, Berkeley, U.S.A)
Xueying Zhang	(Nanjing Normal University, China)

Table of Contents

A Text Mining Approach towards Knowledge Management Applications

Alexandre L. Gonçalves [1], Fabiano Beppler [1,2], Alessandro Bovo [1,2], Vinícius Kern [1,2], Roberto Pacheco [1,2,3]

[1] *Stela Institute, Florianópolis, SC, Brasil {a.l.goncalves, fbeppler, alessandro, kern}@stela.org.br*

[2] *Knowledge Engineering and Management Pos-Graduation, Federal University of Santa Catarina, Florianópolis, SC, Brazil*

[3] *Department of Computing and Statistics, Federal University of Santa Catarina, Florianópolis, SC, Brazil {rpacheco}@inf.ufsc.br*

Abstract

The recognition of entities and their relationships in document collections is an important step towards the discovery of latent knowledge as well as to support knowledge management applications. The challenge lies on how to extract and correlate entities, aiming to answer key knowledge management questions, such as; who works with whom, on which projects, with which customers and on what research areas. The present work proposes a knowledge mining approach supported by information retrieval and text mining tasks in which its core is based on the correlation of textual elements through the LRD (Latent Relation Discovery) method. Our experiments show that LRD outperform better than other correlation methods. Also, we present an application in order to demonstrate the approach over knowledge management scenarios.

1. Introduction

The knowledge has become an important strategic resource for organizations. The generation, codification, management and sharing of the organization knowledge is essential for the innovation process. To know who works with whom, on which projects, with which customers and on what research areas is an important step towards the understanding of intra or extraorganizational relationships.

In the last years the amount of documents has been increased considerable, as much in organizations as in the Web. We state that documents, instead of organizational databases, are the primary resource to reveal latent knowledge, once they keep registered relevant textual patterns (entities), such as, people, organizations and projects, and how such entities are related to each other.

In this work, we present an entity-based knowledge mining approach to support knowledge management tasks. Thus, through the combination of extraction and retrieval of information and text mining tasks, we intend to unveil different levels of connectivity among entities through the projection of collaborative networks or knowledge maps. Such networks are useful tools to provide insights, for instance, about people relationships, that can be spontaneous (i.e., they have common interests and act based on this) or inducted (i.e., they have worked or are working together in a couple of projects).

The rest of the paper is organized as follows. We present the work background in Section 2. Our text mining approach is presented in Section 3. Results are reported in Section 4. Section 5 presents a knowledge management application and finally, we

conclude the paper and discuss future work in Section 6.

Background

Named Entity Recognition (NER) has been applied with success in the identification and classification of textual elements, such as, people, organization, places, monetary values and dates taken into account document collections [Grover et al. 2002], [Cunningham 2002], [Zhu et al. 2005b], [Brin, 1998], [Soderland, 1999], [Ciravegna, 2001]. As result of the process, for each document a set of entities is extracted. Thus, applying co-occurrence based methods the connectivity among entities can be achieved in order to indicate insights toward knowledge discovery.

Co-occurrence methods are important, for instance, in the identification of collocations[1] [Manning and Schütze 1999], information retrieval through vector expansion [Gonçalves et al. 2006] and also as the core for the current proposed. Such methods aim to correlate textual elements aiming to unearth latent relationships.

In this context are t test, chi-square (χ), phi-squared (φ) [Conrad and Utt 1994], [Church and Gale 1991], Z score [Manning and Schütze 1999], [Vechtomova et al. 2003], Mutual Information (MI) [Church and Hanks 1990] or derivations of Mutual Information (VMI) [Vechtomova et al. 2003]. Also, more empirical methods have been applied such as

[1] Natural sequence of words, which possibly, identifies candidate concepts to be extracted from written information.

CORDER [Zhu et al. 2005a] and Latent Relation Discovery (LRD) [Gonçalves et al. 2006].

First of all, the result of the process obtained through the correlation of entities provides direct relationships among entities. However, it is only useful for preliminary analysis on knowledge management applications. Additionally, indirect relationships can also be achieved through clustering algorithms. Such techniques have been used in a wide range of application domains, such as, information retrieval, data mining, machine learning and pattern recognition. They have as main target the grouping of similar objects in the same class [Hair et al. 1998], [Johnson and Wichern 1998], [Halkidi et al. 2001].

All methods and techniques discussed so far are the basis for the proposed approach on achieving knowledge management applications. Knowledge management is seemed as systematic and disciplined actions in which organization can take advantage to get some return [Davenport and Pruzak 1997]. According to Schreiber (2002), knowledge management is an important tool for the enhancement of the organizational knowledge infrastructure. In this scenario, the information technology has an important role in the process of transformation of the knowledge, from tacit to explicit [Marwick 2001]. Thus, we state making explicit entities and their relationships through information extraction and retrieval, and text mining techniques is an important step toward knowledge management applications, such as, communities of practice [Lesser and Storck 2001], [Wenger 1998], expertise location [Marwick 2001] and competency management [Dawson 1991], [Hafeez et al. 2002].

Proposed Approach

The proposed approach (Figure 1) is an extension of the traditional knowledge discovery from textual database model. Traditionally, textual elements are extracted and applied in the data mining phase aiming to reveal useful patterns [Mooney and Nahm 2005]. Our approach is concentrated as much in the extraction of textual elements (i.e., entities and concepts) as in the correlation of such elements. Thus the extraction and correlation of textual elements are the basis for the data mining and information retrieval phases aiming to promote support to knowledge management applications.

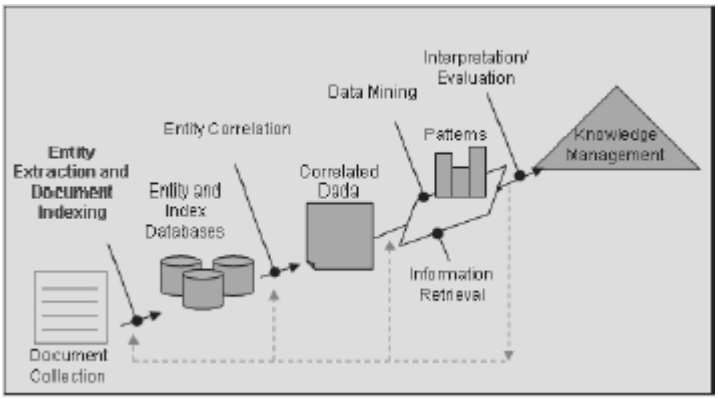

Figure 1: Text Mining approach toward Knowledge Management Applications

Next sections discuss the main phases of the approach, including the extraction and correlation of entities, the composition of the entity database and its use by information retrieval and clustering techniques to support knowledge management applications.

Entity Extraction

The entity extraction phase called Named Entity Recognition (NER) aims to discover proper names, their variations and classes [Cunningham 2002], [Grover et al. 2002]. A named entity (NE) can be defined as a textual element which represents an object in a physical or abstract world. Formally, an entity can be defined as a vector E composed of description, class and additional information, i.e., E={description, class, <additional information>}. Additional information may indicate, for instance, the positions where such patterns occur in the document.

The NER process is mainly composed of two components, that is, lexical structures and patterns. The lexical structures are essential for the process and represent the knowledge base [Guthrie 1996]. Every class taken into account during the NER phase (e.g.: person, organization and project) is associated with a lexical table. Each lexical table stores a set of words that identify itself (e.g.: the person lexical table would have names such as “John” and “Smith”). Additionally, patterns are common in written language and they represent a sequence of words which can be classified (e.g.: an entity similar to “<content> Institute” is classified as “Organization” or “Institute”, or “<content>, Dr” is classified as a “`Person/Doctor”. In this direction, the utilization of regular expressions is an important tool toward the identification of patterns which may be named/classified.

Document Id	Source Entity (SE)	Freq	Target Entity (TE)	Freq	Distance	Partial Relation Strength
1	E1	4	E2	2	2.0000	0.4387
1	E1	4	E3	3	2.0731	0.4938
1	E1	4	E4	1	2.3634	0.3094
1	E1	4	E7	1	2.8540	0.2562
1	E2	2	E3	3	2.3412	0.3123
1	E2	2	E4	1	2.6887	0.1632
1	E2	2	E7	1	3.0805	0.1424
1	E3	3	E4	1	2.2642	0.2584
1	E3	3	E7	1	2.5654	0.2280
1	E4	1	E7	1	3.8074	0.0768
2	E1	2	E3	2	2.0000	0.5850
2	E1	2	E4	1	2.1462	0.4088
2	E1	2	E5	2	2.7925	0.7771
2	E3	2	E4	1	2.6887	0.3263
2	E3	2	E5	2	2.2925	0.9465
2	E4	1	E5	2	3.3877	0.5542
3	E2	3	E6	2	2.0975	0.7827
3	E2	3	E7	2	2.7770	0.3511
3	E6	2	E7	2	3.3877	0.4270

Table 1: List of intra-document weights for each relation among entities with seven entities and three documents

Entity Correlation

Correlation methods have been widely applied to the identification of collocations. Additionally, methods may also be applied to any textual element to indicate proximity, or in our context, the relation strength. To the present work, textual elements are regarded as entities.

Our proposal use the LRD algorithm [Gonçalves et al., 2006] as the method to establish relations among entities taking into account tree aspects: (a) co-occurrence: two entities co-occur if they appear in the same document; (b) distance: the distance of all co-occurrences intra documents is

taken into account; and (c) relation strength: given an entity, E1, the relation strength between two entities E1 and E2 takes into account their co-occurrence, mean distance, and frequency in co-occurred documents as defined in Equation 1. The greater the mean distance is, the smaller the relation strength. Generally, the relation strength between E1 and E2 is asymmetric depending on whether E1 or E2 is the target.

$$R(E1,E2) = \hat{p}(E1,E2) \times \sum_i \left(\frac{f(Freq_i(E1)) \times f(Freq_i(E2))}{m_i(E1,E2)} \right), \quad (1)$$

where $f(Freq_i(E1))=tfidf_i(E1)$, $f(Freq_i(E2))=tfidf_i(E2)$, and $Freq_i(1)$ and $Freq_i(2)$ are the numbers of occurrences of *E1* and *E2* in the i^{th} document, respectively. The term frequency and inverted document frequency measure *tfidf* is defined as $tfidf_i(j)=tf(j)*\log_2(N/df_j)$, where $tf_i(j)=f_i(j)/\max(f_i(k))$ is the frequency $f_i(j)$ of entity *j* in the i^{th} document normalized by the maximum frequency of any entity in the i^{th} document, *N* is the number of documents in the corpus, and df_j is the number of documents that contain the entity *j*.

Composition of the Entity Database

As result of the entity extraction and correlationprocesses, a database with pairs of related entities is created. Each pair, represented by a source entity (SE) and a target entity (TE) is calculated based on LRD method. Given a pair <*SE*, *TE*>, the relation strength is stored as shown in the Table 1. For this example, seven entities extracted from three documents were correlated.

The table is indexed and stored as an inverted index. So given a SE, it is possible to retrieve all related TEs sorted by their relation strength. For instance, the relation strength between the target entity (TE) *E*3 and the source entity (SE) *E*1 is computed using the relative frequency (number of documents in which the relation occur by the total number of documents) times the partial weights summation, i.e., $R(E1,E3) = 2/3 * (0.4938+0.5850) = 0.7192$.

Information Retrieval and Pattern Generation

As mentioned, the entity database is indexed and through information retrieval techniques, answers to knowledge management questions, such as, which the related projects considering a particular SE, makes possible.

Such information is useful in order to produce an initial map about the document collection regarding entities and their relationships. However, it only enables the establishment of direct relationships, i.e., when entities co-occur in the same document. To overcome that, we apply an entity clustering phase in order to obtain more complex relationships. So far, we have applied a simplified but quite fast algorithm, defined as fast clustering to create these maps, called knowledge maps. Given a SE, the k most TEs are retrieved. It will compose all centroids (first level) of the map. For each centroid a new search is carried out up to a threshold (second level). By using Table 1, all entities from a particular cluster are connected, that is, intra-cluster correlation. The process is repeated among entities from different clusters in an inter-cluster correlation process. Before presentation phase, entities which occur in multiple clusters are

merged. In this way, only one entity remains with multiple references for its clusters and other entities.

Evaluation

As the core of the work lies on the entity correlation, the present work proposes a model towards the evaluation of the relation strength among entities. Although, such task tends to be not easy, mainly due to the lack of standard data sets to tackle the correlation of textual element, we intend to compare the precision of LRD and other standard correlation methods.

The main evaluation approaches can be defined as: (a) Quantitative methods judge whether results achieved by the model, based on quantifiable parameters, are suitable. For example, a classic method for analyzing hierarchical agglomerative clustering is the cophenetic correlation coefficient [Sokal and Rohlf, 1962], [Halkidi et al., 2001]. The Square Error Criterion is commonly used to evaluate the efficiency of numerical data clustering [Duda and Hart, 1973]; (b) Gold standard approaches compare the learned model to an “ideal” model produced a priori by domain experts. These are typical in information retrieval, text categorization and information extraction, e.g., MUC [DARPA, 1995], TREC[1], SMART[2] e Reuters[3]. Their primary disadvantage is that standard collections are expensive to produce. Moreover, they are intrinsically subjective since they are based on expert opinion,; and (c) Task oriented evaluations examine algorithms

[1] http://trec.nist.gov/

[2] http://www.dcs.gla.ac.uk/idom/ir_resources/test_collecti ons/

in the context of applications. They are concerned whether the learning algorithm has produced a model that properly works. Tonella et al. [2003] discuss some of the problems associated with such approach including its cost and the need for careful design in order to minimize subjectivity.

Due to the problem context the gold standard and task oriented approaches would be more suitable. However, in general it demands high costs regarding time and people to create datasets as well as it introduces a bias due to the subjectivity. Intending to overcome such limitations, we propose a valuations model discussed below.

Results

In our experiment we have used the LRD method and compared it with other four statistics methods (MI, VMI, Phi-squared and Z score) in the relation strength establishment among entities.

The evaluation is based on a set of 2500 papers from the “Semantic Web” and “Ontology” areas. For each paper the NER process is applied and the result stored as a vector. Each element of the vector represents an entity composed of description, class (Person, Organization and Research Area) and its positions through the document. From the 2301 extracted entities, 970 are organizations, 914 are people and 417 are research areas. In order to avoid subjectivity, the relation between an SE and its TEs is firstly defined by calculating the joint frequency through a traditional search engine.

We state that the joint frequency extracted from documents in the Web which mention the relation R<E1, E2> is an indicative of relatedness. So, given an entity E1, a search engine available on the Web

was used to establish the joint frequency with its related entities. The joint frequency is also an indicative of order or importance. As result, each SE will produce a list with its related TEs and used during the evaluation phase. To analyze the precision, *Spearman's* correlation method was considered (Equation 2).

$$RA = 1 - \frac{6\sum_{i}(R_{i,CM} - R_{i,SE})^2}{N^3 - N}, \qquad (2)$$

where $-1 \leq RA \leq 1$ (1 indicates perfect correlation), $R_{i,CM}$ indicates the order of the entity for a specific correlation method, $R_{i,SE}$ the order obtained through the search engine and N the number of entities used in the query.

For each entity, the top 10 related entities in all correlation methods LRD (*M*1), *Phi-squared* (*M*2), MI (*M*3), VMI (*M*4), and *Z score* (*M*5) were used. Nevertheless, different entities can be selected by a particular method. Generally, it tends to produce lists TEs greater than 10. Also, the study is based on different window sizes, being 50, 100, 200 and no window. Window size is used to validate the relation between two entities, that is, if two entities which co-occur in the same document are out of some specified range, the relation is not valid.

Table 2 presents an example using the "Semantic Web" term (entity in the research area class) and window size of 50. In order to avoid excessive penalties for a particular method which selects TEs beyond the *N* threshold the ranking is normalized. If a specific TE is not selected by some method, the index representing the order is *N*+1. If the

entity order, from *M*1 to *M*5, is equal to *N*+1 and the index generated by search engine is also lesser than *N*+1, or the entity position from *M*1 to *M*5 is different of *N*+1, the partial *Spearman's* index $(R_{iCM} R_{,})^2$ is calculated for the i^{th} pair, otherwise, the correlation value is not taken into account. Best results were achieved by LRD, *Phi-squared* and *Z score*, with 0.930, 0.601 and 0.372, respectively.

Table **3** presents the summarized *Spearman´s* correlation index (regarding all 2301 entities) for *organization*, *person* and *research area* classes as well as the average value. The LRD algorithm achieve the best results followed by the Phi-squared and *Z* score methods, while the MI and VMI methods present the worst results. Among the three classes, the *person* class has the worst performance. The MI and VMI problem is due to theirs deficiency to deal with low frequencies. In this case high values are generally attributed to relations with little importance.

Table 2 (Next page): Establishment of the Spearman's correlation for the "Semantic Web" entity and its most related pair taken into account 5 methods, where SE=search engine, order=order established via SE, from *M*1 to *M*5 the order established by correlation methods (LRD, Phi-squared, MI, VMI and *Z* score) and from *R*1 to *R*5 the Spearman's partial value $(R_{i,CM} - R_{i,SE})^2$ for each correlation methods.

Related entities	SE	Order	M1	M2	M3	M4	M5	R1	R2	R3	R4	R5
Xml	3.240.000	1	6	11	11	11	11	25	100	100	100	100
Rdf	3.140.000	2	2	11	11	11	11	0	81	81	81	81
ontology	1.530.000	3	1	11	11	11	11	4	64	64	64	64
networks	1.470.000	4	10	3	11	11	11	36	1	49	49	49
web services	1.460.000	5	4	11	11	11	11	1	36	36	36	36
Owl	732.000	6	8	11	11	11	11	4	25	25	25	25
knowledge management	713.000	7	11	6	11	11	6	16	1	16	16	1
agent	623.000	8	5	11	11	11	11	9	9	9	9	9
interoperability	547.000	9	7	1	11	11	11	4	64	4	4	4
information systems	474.000	10	11	8	11	11	11	1	4	1	1	1
environments	471.000	11	11	4	2	4	11	*	49	81	49	*
reasoning	439.000	12	9	11	11	11	11	9	*	*	*	*
patterns	401.000	13	11	9	3	10	1	*	16	100	9	144
Daml	302.000	14	3	11	11	11	11	121	*	*	*	*
User interface	262.000	15	11	11	11	6	2	*	*	*	81	169
simulation	253.000	16	11	10	5	11	8	*	36	121	*	64
knowledge representation	237.000	17	11	7	11	11	10	*	100	*	*	49
hypertext	231.000	18	11	11	10	5	3	*	*	64	169	225
intelligent systems	164.000	19	11	11	8	7	11	*	*	121	144	*
Trees	157.000	20	11	11	11	9	7	*	*	*	121	169
electronic commerce	140.000	21	11	2	4	11	9	*	361	289	*	144
hypermedia	123.000	22	11	11	11	8	11	*	*	*	196	*
problem solving	118.000	23	11	11	6	11	5	*	*	289	*	324
relational database	83.500	24	11	5	1	2	4	*	361	529	484	400
database management	75.600	25	11	11	11	3	11	*	*	*	484	*
ontology engineering	68.000	26	11	11	9	1	11	*	*	289	625	*
system architecture	44.300	27	11	11	7	11	11	*	*	400	*	*
								230	1308	2668	2747	2058
							Spearman	0,930	0,601	0,186	0,161	0,372

Spearman [-1,1]		*Organization*	*Person*	*Research Area*	*Average*
No window	LRD	0.4231	0.4496	0.4236	0.3979
	Phi-squared	0.1487	-0.0291	0.0834	0.1306
	MI	0.0797	-0.1525	-0.0071	0.0515
	VMI	0.0797	-0.1525	-0.0071	0.0515
	Z Score	0.1487	-0.0291	0.0834	0.1306
Window (50)	LRD	0.3286	0.2572	0.2866	0.2739
	Phi-squared	0.0598	-0.0562	-0.0048	-0.0180
	MI	0.0157	-0.1715	-0.0700	-0.0542
	VMI	0.0768	-0.1367	-0.0142	0.0173
	Z Score	0.1126	-0.0764	0.0198	0.0231
Window (100)	LRD	0.3423	0.2788	0.3242	0.3515
	Phi-squared	0.1073	-0.0519	0.0488	0.0910
	MI	0.0624	-0.1551	-0.0196	0.0339
	VMI	0.0990	-0.1344	0.0151	0.0808
	Z Score	0.1161	-0.0884	0.0345	0.0759
Window (200)	LRD	0.3847	0.3452	0.3759	0.3980
	Phi-squared	0.1380	-0.0231	0.0923	0.1620
	MI	0.0803	-0.1298	0.0111	0.0827
	VMI	0.0824	-0.1317	0.0276	0.1320
	Z Score	0.1075	-0.0844	0.0549	0.1417

Table 3: Spearman values between -1 and 1 for the *organization*, *person* and *research area* classes as well as the average value for different window configurations

Knowledge Management Scenarios

Tools for analysis of entities and their relationships represent important resource towards knowledge management applications, such as, communities of practice, social networks, expertise location and competency management.

Such applications have in common the use of entity relations to express different objectives. While communities of practice are broader, that is, all entity classes can be projected together, social networks have its focus on person class. Communities of

practice are therefore a more general class of application, being easily configured to achieve social networks.

According to Alani et al. [2003], communities of practice are established by groups whose members are interested in a particular job, procedure, or work domain. Lesser and Stoch [2001] define communities as groups engaged in sharing and learning, based on common interests. On the other hand, organizations are more and more aware about the knowledge and skills of collaborators as their most valuable resource. Know who knows what is now a critical activity. Equally important is to know who knows whom, both inside and outside organizations. Thus, such networking identification may be useful as much in the optimization of project resources as in the creation of new business opportunities.

Similarly to communities of practice, expertise location and competency management, have an important role inside organizations, mainly aiming the development of core competencies. According to Dawson [1991], core competencies are the combination of learnt skills or under development. It can foster or help on the establishment of business strategies [Hafeez et al., 2002]. So, methods used in the discovery and evaluation of core competencies have shown to be useful in a couple of activities toward the management of the organizational intellectual capital.

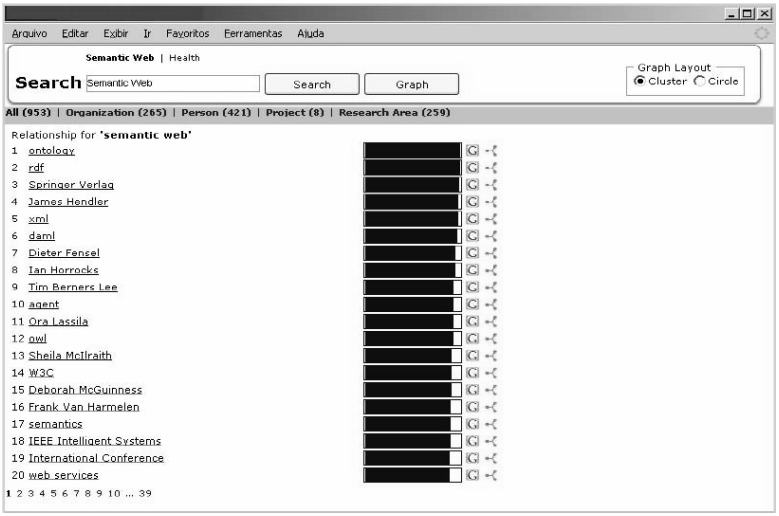

Figure 2: Entities and their relationships regarding all classes

Aiming to explore and provide ways to understand entities and their relationships we have developed some tools. The Figure 2 shows a tool in which by informing a SE, the most related TEs are retrieved taken into account some classes (e.g.: *organization*, *person*, *project* and *research area*) their relation strength. Also, the tool makes possible the analysis of the most relevant relations for each class, and promotes an easy way to inspect competencies or even to retrieve experts regarding a particular subject (*research area*).

Despite its utility such presentation can be enhanced by showing the same information graphically. We have applied a clustering algorithm to unveil latent relations and to group close entities. In order to facilitate the visualization, different classes are identified by different colors and shapes (squares

represent main entities, the cluster centroids, whist circles represent related entities).

Thus, through these representation ways (textual or graph) it is intended to provide insights to help on managing the organizational knowledge.

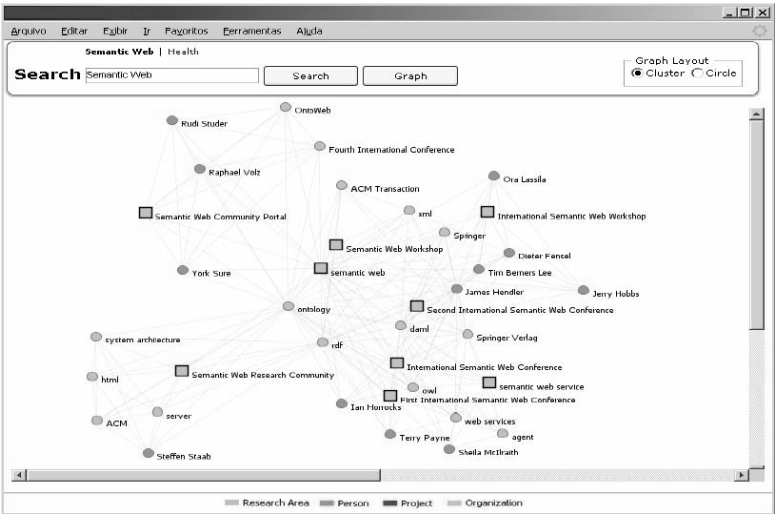

Figure 3: Clusters of entities and their relationships

Conclusion and Future Work

We have shown a text mining approach based on co-occurrence in order to establish the relation strength among textual elements (entities, concepts). As result of the process databases representing documents and entities are created and indexed. Thus, through clustering and information retrieval tasks, we intend to unveil latent knowledge and support the decision making process on knowledge management area. Also, we have compared the *Spearman's* correlation between a search engine and other five methods. LRD has shown the best results

taken into account different entity classes and window sizes.

Our future work is three-fold. First, we are working on refining proposed method aiming to improve metrics used to establish the relation strength among entities as well as to improve the clustering method. Second, benchmarks with other architectures, similar we have presented here, are intended. Finally, entities and their relations constitute primary resource for network analysis and ontologies. In this sense, improvements on knowledge management applications as those shown here are on the way.

References

Alani, H., Dasmahapatra, S., O'hara, K. and Shadbolt, N. (2003) “Identifying communities of practice through ontology network analysis”, IEEE Intelligent Systems, v. 18, n. 2, p. 18-25.

Bontcheva, K., Maynard, D., Tablan, V. and Cunningham, H. (2003) “GATE: A Unicode-based Infrastructure Supporting Multilingual Information Extraction”, In *Workshop on Information Extraction for Slavonic and other Central and Eastern European Languages*. Held in conjunction with the 4th International Conference "Recent Advances in Natural Language Processing" (RANLP'2003), Bulgaria.

Brin, S. (1998) “Extracting Patterns and Relations from the World Wide Web”, In *Proceedings of WebDB* (1998), pages 172-183.

Church, K. and Gale, W. (1991) “Concordances for parallel text”, In *Proceedings of the Seventh Annual Conference of the University of Waterloo Centre for the New OED and Text Research*, pages 40-62.

Church, K. and Hanks, P. (1990) “Word association norms, mutual information, and lexicography”, Computational Linguistics, v. 16, n. 1, p. 22-29.

Ciravegna, F. (2001) “Adaptive Information Extraction from Text by Rule Induction and Generalisation”, In *Proceedings of IJCAI* (2001).

Conrad, J. G. and Utt, M. H. (1994) “A System for Discovering Relationships by Feature Extraction from Text Databases”, SIGIR, p. 260-270.

Cunningham, H. Gate (2001) “A General Architecture for Text Engineering”, Computers and the Humanities, v. 36, n. 2, p. 223-254.

DARPA (Defense Advanced Research Projects Agency) (1995), In *Proceedings of the Sixth Message Understanding Conference*, Morgan Kaufmann.

Davenport, T. H. and Prusak, L. (1997) “Information ecology: Mastering the information and knowledge environment”, Oxford University Press.

Dawson, K. (1991) “Core competency management in R&D organizations”, In *Technology Management: The New International Language*, Dundar Kocaoglu and Kiyoshi Niwa (eds.), New York, Institute of Electrical and Electronics Engineers, p. 145-148.

Duda, R. and Hart, P. (1973) “Pattern classification and scene analysis”, Wiley, New York.

Gonçalves, A. L., Zhu, J., Song, D., Uren, V. and Pacheco, R. (2006) “LRD: Latent Relation Discovery for Vector Space Expansion and Information Retrieval”, In *Proceedings of the 7th International Conference on Web-Age Information Management* (WAIM 2006), J.X. Yu, M. Kitsuregawa, and H.V. Leong (Eds.): Lecture Notes in Computer Science (LNCS), Hong Kong, China, p. 122-133.

Grover, C., Gearailt, D. N., Karkaletsis, V., Farmakiotou, D., Pazienza, M. T. and Vindigni, M. (2002) “Multilingual XML-Based Named Entity Recognition for E-Retail Domains”, In *Proceedings of the 3rd International Conference on Language Resources and Evaluation* (LREC 2002), Las Palmas, pages 1060-1067.

Guthrie, L., Pustejowsky, J., Wilks, Y. and Slator, B. M. (1996) “The Role of Lexicons in Natural Language Processing**”,** Communications of the ACM, v. 39, n. 1, p. 63-72.

Hafeez, K., Zhang, Y. and Malak, N. (2002) “Identifying core competence”, IEEE Potencials, v. 49, n. 1, p. 2-8.

Hair Jr., J. F., Anderson, R. E., Tatham, R. L. and Black, W. C. (1998) “Multivariate data analysis”. Prentice-Hall, Upper Saddle River, 5. ed., New Jersey.

Halkidi, M., Batistakis, Y. and Vazirgiannis, M. (2001) “On clustering validation techniques”,

Journal of Intelligent Information Systems, v. 17, n. 2-3, p. 107-145. Johnson, R. A. and Wichern, D. W. (1998) Applied multivariate statistical analysis, New Jersey: Prentice-Hall, 4th edition.

Lesser, E. L. and Storck, J. (2001) “Communities of practice and organizational performance”, IBM Systems Journal, v. 40, n. 4, p. 831-841.

Manning, C. D. and Schütze, H. (1999), Foundations of statistical natural language processing, The MIT Press, Cambridge, Massachusetts.

Marwick, A.D. (2001) “Knowledge management technology”. IBM Systems Journal, v. 40, n. 4, p. 814-830.

Mooney, R. J. and Nahm, Un Y. (2005) “Text Mining with Information Extraction”. In: Proceedings of the International MIDP colloquium, September 2003, Bloemfontein, South Africa, Daelemans, W., du Plessis, T., Snyman, C. and Teck, L. (Eds.), Van Schaik Pub., South Africa, p. 141-160, 2005.

Schreiber, G., Akkermans, H., Anjewierden, A., Hoog, R. de, Shadbolt, N., Velde, W. V. de and Wielinga, B. (2002), Knowledge engineering and management: The CommomKADS Metlhodology, The MIT Press, 3rd edition.

Soderland, S. (1999) “Learning Information Extraction Rules for Semi-Structured and Free Text. Machine Learning”, n. 34, v. 1, p. 233–272.

Sokal, R. R. and Rohlf, F. J. (1962) “The Comparison of Dendrograms by Objective Methods”, TAXON, v. 11, p. 33-40.

Tonella, P., Ricca, F., Pianta, E., Girardi, C., Di Lucca, G., Fasolino, A. R. and Tramontana, P. (2003) “Evaluation Methods for Web Application Clustering”, In *Proceedings of the 5th International Workshop on Web Site Evolution*.

Vechtomova, O., Robertson, S. and Jones, S. (2003) “Query expansion with long-span collocates”, Information Retrieval, v. 6, n. 2, p. 251-273.

Wenger E. (1998), Communities of practice, learning meaning and identity, Cambridge University Press, Cambridge, MA.

Zhu, J., Gonçalves, A., Uren, V., Motta, E. and Pacheco, R. (2005a) “Mining Web Data for Competency Management”. In

Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence (WI 2005), France.

Zhu, J., Uren, V. and Motta, E. (2005b) "ESpotter: Adaptive Named Entity Recognition for Web Browsing", In: *Proceedings of the 3rd Conference on Professional Knowledge Management*WM2005), Kaiserslautern, Germany.

Data Integration in Current Research Information Systems

Maximilian Stempfhuber[1]
[1]GESIS / Social Science Information Centre (IZ), Bonn, Germany
Stempfhuber@iz-soz.de

Abstract.

Information portals currently follow the principle of aggregation – connecting previously unconnected collections of information to give users a single place of access to this information. The paper argues that aggregation alone will not guaranty the level of quality of service which scientific users demand. Starting with findings from recent user surveys, essential features of information portals are presented and the problems and shortcomings of current solutions are discussed. The second part of the paper presents a model for scientific information portals which focuses on integration, which means that information is not only collected but semantic differences between individual offerings are treated in a coordinated way. As an example for transferring the model to real world information portals, the new social science portal SOWIPORT is used.

Introduction

Looking at the landscape of scientific information, many different offerings can be found. In most cases they are provided by libraries, information centres, research institutes and commercial information providers, which make their results, services or the materials they collected publicly available. Besides this, the new publishing paradigm

of the World Wide Web allows every single user to simultaneously collect and redistribute information with only a very low barrier concerning costs and technology. This leads to a polycentric information landscape (see figure 1) – without a central institution organizing collaboration and workflow – and a fragmentation of information and information services, mostly due to organizational and domain-specific aspects, personal preferences and interests, and available resources.

Challenges in a changing information landscape

Up to only a few years ago it was the task of traditional information providers (publishers with their print publications, libraries with their catalogues and information centres with their reference databases) to coordinate access to information and guarantee a homogeneous level of service on the bases of sophisticated standards (right part of figure 1). But with the advent of electronic publishing and reduced entry barriers in the market of scientific information (Cigan 2002), researchers and publishers make use of the new technologies and offer complementary services to the public. Library catalogues and reference databases now are only additional modules in a worldwide and connected portfolio of information offerings, where standards are no longer applied. Most evidently, the loss of standardization is seen when looking at the methods of content indexing applied by different players (M1 to M6 in figure 1). Besides reference databases with a full set of metadata, abstract and keywords from a controlled, domain-specific vocabulary – but often without the full text document - (M1) nearly every type of content

indexing can be found down to the full text document put on a researchers website without any further bibliographic information and indexed only by search engines (M5).

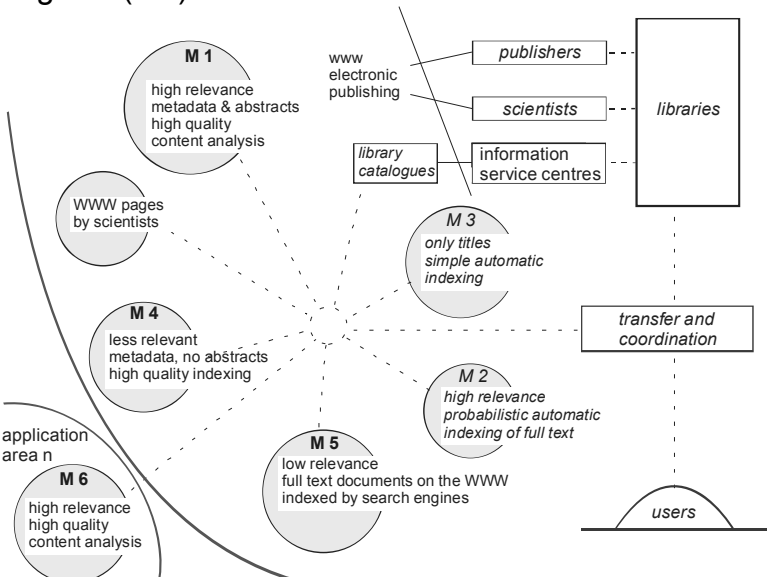

Figure 1: The change from a centralistic to a polycentric information landscape (Krause 2006, modified)

In addition, primary and secondary (empirical) data, literature references, research projects, internet resources and information about the structure of a scientific domain (e.g. journals, institutes and experts) are distributed across many offerings and sites, different methodologies for organizing and presenting this offerings are applied (e.g. for structuring information and for content indexing), and different technologies and access modes are used.

For the user, this situation has a number of implications. To access all materials relevant to his information need he has to locate these single

offerings, has to assess the quality of the materials they offer, and has to cope with the way information and services are presented: Different user interfaces, metadata schemas, indexing vocabularies, query languages, languages in which the content is presented, and legal access restrictions. Theories and models from information science suggest other, more user oriented ways of designing and building information systems, but many times practical aspects – or missing awareness on the side of information providers – prevent improvements in usability and quality of service.

A number of surveys for (Mann 2006, Poll 2004, RSLG 2002a & 2002b) tried to assess user satisfaction with scientific information services in Germany and in the UK during the last few years. They now support the theories and models from information science and give a clearer picture of what users really want:

Domain-specific organization: Users demand for domain-specific organization of information, so they have a single point of access to all potentially relevant materials from their discipline. With it goes extensive and precise content indexing – especially in cases where only metadata or references to the primary information is available – allowing them to search for specific semantic features of documents or survey data at a very detailed level. In the social sciences, this can be realized by using domain-specific thesauri and classifications, and by additionally describing the content with e.g. the scientific methods used or the time frame or geographic region under study.

Interdisciplinary connections: At the same time, users see the need for crossing the borders between domains and disciplines. Since more and more research is carried out at the intersection of multiple disciplines, it is essential that the access to information accommodates for these complex information needs. Given the fact, that on average a researcher spends only half a day per week for information procurement[1], information systems have to support this type of searching in a way the burden on the user is minimized and queries to different data collections can be carried out efficiently. Similarly, connections between disciplines allow users who can not determine the most relevant domain for their information need to start their search on a more general level and then drill down to the discipline or information source suited which suits them best.

Broad scope of information: To satisfy most information needs, all relevant types of information should be made accessible at one place: Primary data (e.g. surveys and other raw data), secondary data (e.g. aggregate and time series data), and references to literature, research projects, internet resources, experts, networks, software etc. The benefit for the user is maximized if all different types of content are accessible in an integrated (and interlinked) way and indexed semantically rich so that users may seamlessly switch between the different modes of access of these heterogeneous data.

[1] Friedlander 2002 (table 15) states that around 16 hours per week are used for obtaining, reviewing, and analyzing information from all sources to support both teaching and research, with the highest numbers in the social sciences.

Quality of service: To help users satisfy their information needs, results should be limited to a manageable amount of relevant information (no information overflow), noise and low quality content should be avoided and all materials should be instantly accessible right from the desktop ("now-or-never" paradigm) or they might not be used at all. This implies extending – or even replacing – the Boolean retrieval model often used with scientific information by alternative models which allow for relevance ranked results, and bringing content into a digital form so that it can directly be downloaded and used on the researcher's computer.

Search engines are not enough: Besides services from libraries and other institutional information providers, search engines are used for a number of reasons. Most likely, the very simple user interface (in comparison with that of OPACS, Online Public Access Library catalogues), the relevance ranking of the results and the direct access to the materials found add to the popularity of these general – not domain-specific – search engines. Interestingly, older researchers seem to be less critical about what can be found with search engines than younger researchers which tend to stress the sometimes low quality and incompleteness of the results. Google Scholar[1] is an effort to restrict the search space of general search engines to scientific documents by acquiring scientific content from publishers, hoping to get more relevant results and less noise.

Informal communication: Besides the need for accessing information for carrying out research is also a strong need for communication, like discussing

[1] See http://scholar.google.com

results with colleagues in discussion boards or exchanging information and papers via e-mail. This informal communication (in contrast to the formal communication by means of classical publications) is of more dynamic nature, subject to small and even closed groups of researches (invisible colleges), and currently often separated from the information services offered to a discipline. The growing importance of this type of interaction for the generation of new ideas and even of new types of research and publications (Harnad 1991, Nentwich 2003) makes it inevitable to integrate it more seamlessly with traditional offerings.

To satisfy the formulated user needs, many activities already have been started aiming at the aggregation of content within information portals. Starting with the Open Archives Initiative[1] (OAI) where initially data archives – and nowadays also libraries of universities and institutes – supply metadata on the materials they have stored locally, a culture of sharing and (re-) distributing content of different types has grown. International federated information gateways, like RENARDUS[2], exchange metadata to make their offerings more visible and more easily findable to national and international users. And on the national level, infrastructures like the German-based vascoda[3] portal allow searching and accessing the databases, library catalogues and subject gateways of over 30 German scientific information providers and libraries.

But still, central problems with great influence on the user experience with information services have not been tackled. Despite existing – and often

1 See http://www.openarchives.org
2 See http://www.renardus.org
3 See http://www.vascoda.de

adhered to – standards for user interface design, switching between offerings most of the time involves learning to efficiently interact with the system from scratch. Neither interaction models nor menu structures or feature sets are matched, and the more systems or offerings are connected, the smaller is the number of common features which can be realized at the level of an information portal – a common gateway to distinct services and products which share enough similarities so that users see them as a whole.

Looking at the content level, we face nearly the same situation. Content harvested from or ingested by individual information providers is combined – aggregated – in information portals for searching and browsing. Most effort is currently spent on syntactic and structural aspects, standardizing sets of elements by which content can be described to allow field-based searching (e.g. the Dublin Core Metadata Element Set[1], DC, or the Data Documentation Initiative[2], DDI). What currently are not sufficiently handled are the differences in the indexing vocabularies used for filling in the metadata elements. Since many information providers use their own vocabulary (e.g. a specialized thesaurus or classification), a user's query (i.e. the keywords used for searching) can not easily be matched with the variety of keywords used to describe the same semantic concept in all the different metadata collections aggregated in the portal. This leads to imprecise results, leaving out relevant materials and possibly containing unwanted information.

1 See http://www.dublincore.org/

2 See https://www.icpsr.umich.edu/DDI/index.html

At the same time, traditional information providers, especially libraries and information centres, face a growing competition by internet search engines with their easy to use user interfaces, relevance ranking, and direct linking to electronic resources and are simultaneously challenged by shrinking – or at least not growing – budgets. Models proposed for strengthening the position of scholarly information services are normally a reduction of complexity by simplifying workflow, metadata creation or context indexing (Calhoun 2006; for a critique to this position see Mann 2006), sometimes neglecting that only this assets would allow for advanced retrieval services in the absence of full texts or original data.

In short, many problems with information portals arise from aggregating heterogeneous content at different levels (e.g. structural, semantic and interaction level) without integrating it, which in turn is a prerequisite for high user satisfaction. The remainder of this paper will therefore focus on the semantic level of content integration – the level of heterogeneous vocabularies (e.g. thesauri and classifications) – and will present a model for integrating different types of information and different types of scholarly communication at the user level.

Dealing with heterogeneity in information systems

Heterogeneity can nowadays be seen as one of the central problems – or research challenges – when building integrated information systems. It arises from differences in data structures and content analysis between data collections as soon as these collections are brought together within one information system. At the user's side, the differences sometimes can be

hidden by using simple search functions (e.g. the famous “Google-like” search) and full text indexing, omitting the richness and expressiveness the underlying data would allow for. But as soon as more details of the underlying structures have to be exposed (e.g. keyword search with thesaurus support or a field based search at a more detailed level) the user is faced with fields which are not valid for all databases or with only a basic index instead of a thesaurus (or no help for query formulation at all), because not all databases are indexed with a thesaurus – and those who are use different ones. Standardization might seem as a natural way of dealing with this heterogeneity.

But there are problems with standardization. In many areas where standardization is important for reasons of security, interchange ability, economic gain or technologic advancement, it is also realized that alternatives are needed to make the necessary progress. This is especially the case if regulations can not be imposed on all parties (e.g. information providers in different countries) or if the adaptation of standards is too slow or costly (e.g. in situations where alternative standards have already been applied). The solution may be to open the process of standardization to the idea that there always will be details which can not be standardized and that this fact should be kept in mind and accounted for right from the beginning of the standardization process. The German standardization body, DIN, adapted this strategy in its recent position papers (DIN 2003a, DIN 2003b) about standardization:

“The classic approach to standardization, to achieve compatibility and interoperability by technical

uniformity, reaches its limits where regional or industry-specific solutions have been implemented with great effort (e.g. infrastructures) and subsequently global interoperability has to be assured as a result of globalization or a changed business situation.

SICT recommends regarding standardization also from the viewpoint of providing 'interoperability for existing heterogeneity'.

This task consists of finding a reasonable balance between the desired scope of standardization and the remaining heterogeneity treatment. The costs and the quality losses possibly resulting from the heterogeneity treatment should be seen in relation to the expenditure and the chances of success of further intensified standardization." (DIN 2003a:7)

With this model in mind – to use standards wherever possible but develop alternative means for homogenization right from the start – the interdisciplinary information portal infoconnex[1] has been developed. Infoconnex integrates the German reference databases for pedagogic (FIS Bildung), social sciences (SOLIS) and psychology (PSYNDEX) under one roof and enables the user to cross-search the databases by using any of the three thesauri involved for context indexing of the databases.

The methods used are cross-concordances – bilateral, pair wise mappings between the three thesauri – which are used to transform a user's keyword query to every of the single thesauri. Cross-concordances are intellectually created relations between the terms of each thesaurus, connecting one start term with one or a combination of several target

[1] See http://www.infoconnex.de

terms. Figure 2 shows the underlying principles of cross-concordances and alternative methods for term transfer.

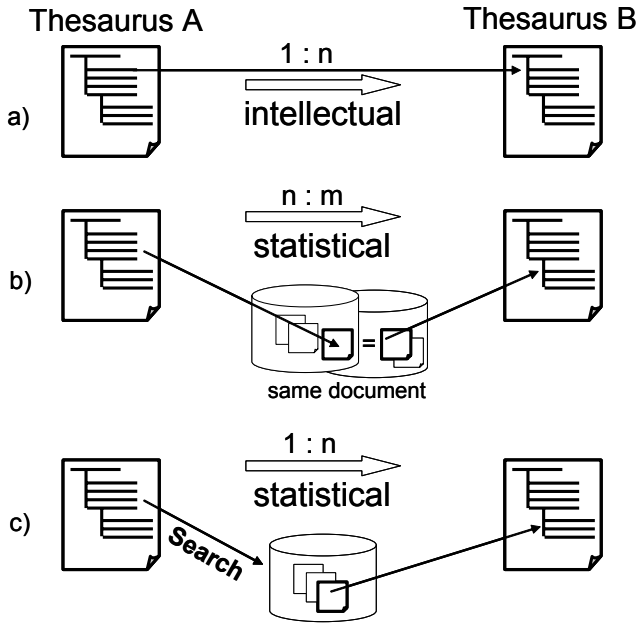

Figure 2: Methods for dealing with heterogeneity between indexing languages

Cross-concordances connect one term of the source vocabulary (thesaurus A in figure 2a) with one or (a combination of) more terms of the target vocabulary (thesaurus B) by the relationships equivalence, similarity or broader and narrower term. The relationships are defined by intellectually analyzing both vocabularies and are checked concerning their retrieval quality by using them for database queries and rating the relevance of the result as high, medium or low.

Two alternative approaches to vocabulary switching are based on statistical analysis of the co-occurrence of terms of both thesauri for a single document. They require no intellectual input and reflect the actual use of the vocabularies for content indexing in the collections used for analysis. Ideally, the same document can be identified in two databases which respectively use one of the two thesauri under question (figure 2b). These two databases can be seen as parallel corpora, holding the same documents with the same content but expressing this content semantically different by using two different vocabularies. In practice, such corpora can be found (sometimes a single database or catalogue uses two different vocabularies for indexing) or they can be constructed by e.g. identifying identical pairs of documents in library catalogues (using a broad and interdisciplinary vocabulary) and reference databases (using a smaller domain-specific thesaurus). In contrast to cross-concordances, which look at individual terms, the statistical approach relates groups of terms from each thesaurus whose likelihood to be used together for describing the same semantic concepts exceeds a certain cut-off value. If no parallel corpora are available, one can simulate them by using terms from one vocabulary to find relevant documents (ranked by relevance) and then statistically relate the keywords of these documents to the terms used for searching (figure 2c).

It is important to notice that the approach based on cross-concordances is conceptually different from the statistical approaches in that former maps between vocabularies without taking the actual use of the vocabularies for content indexing into account,

whereas the later is only based on the actual vocabulary use for indexing. The sets of terms mapped by cross-concordances normally are comprehensible for a user, while the sets of terms generated by statistics may sometimes look farfetched or even unrelated by themselves, but they reflect the frequent co-occurrence of semantic concepts in actual documents.

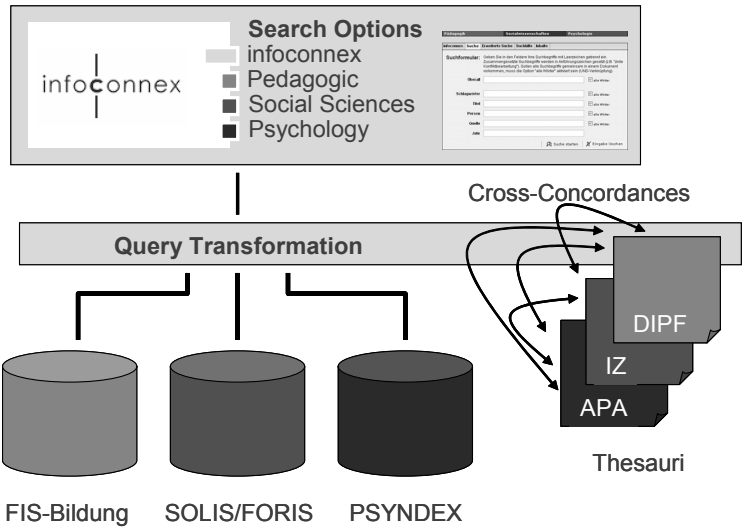

Figure 3: Treating heterogeneity with cross-concordances

The infoconnex portal currently uses cross-concordances to map between three thesauri allowing users to formulate interdisciplinary queries. The user may select keywords from any of the thesauri and the system will automatically determine the appropriate starting thesauri and generate mapped terms for any target database. By using only equivalence relations in a first step of the query process, no additional user interaction is necessary. Only after reviewing the first results, further decisions to expand or narrow the

search are required. This puts the modules for treatment of semantic heterogeneity into action in a way totally transparent to the user, so he can focus on phrasing his information need without dealing with the complexity of the underlying transformations.

To allow advanced users to influence the search at a fine grained level and to visualize the influence the individual mappings between thesauri have at the result set, a graphical user interface based on visual formalisms has been developed (Stempfhuber 2003). Figure 4 shows the basic visualization, a table, used for displaying the structural heterogeneity in an information system.

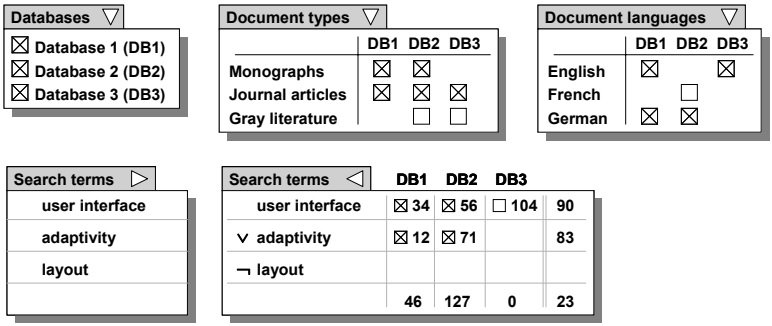

Figure 4: Visualizing heterogeneity in information systems

The top row of tables shows filters, by which a user may select databases and then restrict its search to specific document types or languages. The nature of visual formalisms (Nardi&Zarmer 1993) allows the user to simultaneously get an impression of the semantics in the underlying data (which databases contain what types of documents or documents in which languages) and at the same time interact with the data. In the case of an information system he might deselect certain document types or languages.

The same visualization can be used to let the user decide which vocabulary transformations are used for the individual databases or which of his search terms will be used in which databases. The lower row of tables in figure 4 shows a tabular entry field for search terms in compressed (left) and extended (right) view. Additional “fly-over” windows can also show the actual transformation for each search term and each database. This type of visualization has already been tested for literature databases and for geographic information and has proven its applicability in cross-cultural settings (Stempfhuber et al. 2003).

Integrating information in the social sciences

The methods for dealing with structural and semantic heterogeneity at the user interface level and at the retrieval level build the conceptual basis of a new social science portal, SOWIPORT, which is currently being built by the Social Science Information Centre in Bonn, Germany, and several partner institutes and libraries. The goal of SOWIPORT is to integrate all relevant types of scientific information without losing the expressiveness of the data through standardization: Primary and secondary data, literature references, online journals, scholarly discussion etc. This requires intensive work on the knowledge organization level to integrate heterogeneous indexing vocabularies and data structures. A paramount goal is also to bridge the boarders between different types of information so that users experience the collected data as a ‘whole’, allowing to search and navigate to a specific piece of information along different routes.

SOWIPORT has to been seen in the context of the currently reorganized structure of scientific information in Germany. Figure 5 shows the hierarchical – or cascading – organization of disciplines and portals. The top level portal vascoda represents the central and interdisciplinary entry point to scientific information. Vascoda lets the user search across all disciplines or by selecting a cluster of disciplines, like the cluster Law, Economics & Social Sciences. The search is carried out across over 30 information providers (mostly domain-specific virtual libraries, information centres and commercial hosts) which classify their content according to a classification of disciplines.

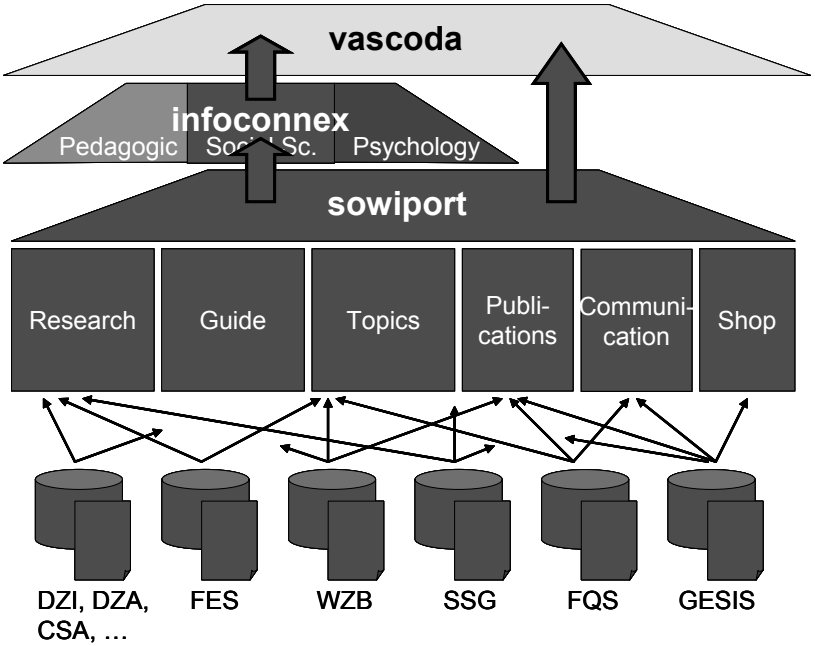

Figure 5: Cascading structure of scientific information offerings

At the level of the disciplines, some of them have closer connections than others, being candidates for a tighter integration which then facilitates interdisciplinary search and research. Infoconnex is a prototype of this kind of intermediary structure. The portal uses modules for the treatment of heterogeneity between the indexing vocabularies of the disciplines and supports users during cross-database (and discipline) search.

At the level of a single discipline, SOWIPORT takes on responsibility to integrate (heterogeneous) information in the social sciences. The portal will deliver this information directly to users from the social sciences as it will also provide the information at the cluster level (infoconnex) and at the top level of the vascoda portal. This cascading structure with divided responsibility (interdisciplinary – related disciplines – single discipline) reduces complexity at higher levels – sacrificing a certain level of detail – but allows for specific treatment of content heterogeneity and user needs at lower levels.

The content of the large number of information providers in SOWIPORT will be organized in several areas, like research (primary and secondary data, references to literature and research projects etc.), a guide to the structure of social sciences (institutes, experts, networks, journals, conferences etc.), topics (e.g. qualitative research, migration etc.), publications (online journals, books, newsletters etc.) and communication (mailing lists, discussion boards etc.).

Aggregating social science information

To collect a critical mass of information that will attract users, different policies have been used. The information available at the Social Science Information

Centre, at its sister institutes within German Social Science Infrastructure Services[1] (GESIS) and at its institutional partners in the Social Sciences Virtual Library[2] (ViBSoz) form a nucleus of well-known and tightly integrated information offerings.

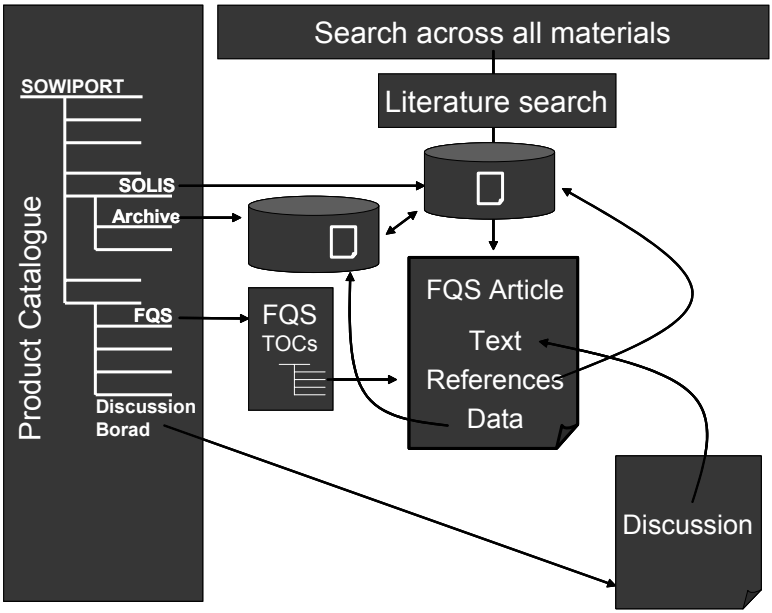

Figure 6: Integration across media

This nucleus is extended with external offerings. Two main policies are being followed, namely the integration of Open Access materials and the negotiation of national licences. In the area of Open Access, SOWIPORT will not only integrate three international journals but additionally materials from the social sciences will be harvested and a repository will be offered to allow researchers to deposit their

[1] See http://www.gesis.org
[2] See http://www.vibsoz.de

publications. Concerning commercial offerings, we were able to get funding from Deutsche Forschungsgemeinschaft[1] (DFG) for 11 volumes (1999-2009) of each of six international databases: CSA Sociological Abstracts, CSA Social Services Abstracts, PAIS International, CSA Worldwide Political Science Abstracts, ASSIA: Applied Social Sciences Index and Abstracts, and Physical Education Index. These national licences entitle researchers with residence in Germany to access the licensed content free of cost. They complement the collections of the SOWIPORT partners at the international level.

Integration of information in SOWIPORT

The content aggregated in SOWIPORT will be integrated not only at the semantic level – by treating the structural and semantic heterogeneity and organizing the content according to the quality of content indexing and integration (Krause 2006) – but also by bridging the boarders between different types of information and between formal and informal communication. Figure 6 gives an impression of how information will be accessible via searching and browsing using an online article of the open access journal FQS[2] as an example. By using the general or the literature-specific search one will find a reference to an FQS article in the literature database SOLIS and follow the link to the online version of the article. Alternatively the article is accessible from SOWIPORTS's product catalogue following the route from publications to FQS's table of contents and on to the article. Looking at the online article, links will be

[1] See http://www.dfg.de

[2] See http://www.qualitative-research.net/fqs/fqs-eng.htm

created from the reference section back to the literature databases and to the data archive, should there be any primary data available. The primary data, of cause could also be found by going to the archive section of SOWIPORT. To bridge between formal an informal communication, online discussion will be integrated into the online journal without the break in media currently found whenever discussion is handled in a separate discussion board and not as an integral part of an online publication.

Last but not least, we will try to transfer experiences with the integration of time-series data and literature made in the domain of market research (Stempfhuber et al. 2002) to the social sciences, supporting users to find related raw data and research activities (projects) and results (literature).

Conclusion

In this paper we argued that in many cases information portals aggregate content relevant to their user group. To cope with the heterogeneity of the content standardization is used which trades specific access to a single product for common access to all products. While aggregation of content is a prerequisite for building adequate information services, it falls short of handling areas which can not be standardized or to bring together different media to form new services and products. By the example of SOWIPORT we presented a model which integrates all types of information identified as relevant for a social science information portal. The model is based on modules for treating semantic and structural heterogeneity in information systems at different levels and connects different media, especially formal and

informal communication to yield new – integrated – information services for the social sciences.

References

Calhoun, K. (2006): The Changing Nature of the Catalog and its Integration with Other Discovery Tools. Final Report. March 17, 2006. Prepared for the Library of Congress by Karen Calhoun (http://www.loc.gov/catdir/calhoun-report-final.pdf).

Cigan, H. (2002): Der Beitrag des Internet für den Fortschritt und das Wachstum in Deutschland. Hamburg Institute of International Economics, Report 217, Hamburg.

DIN (2003a): Standardization in Information and Communication Technology (ICT). German Positions. DIN Deutsches Institut für Normung e.V. Strategy Committee on Standardization in Information and Communication Technology (SICT) (http://www.ni.din.de/sixcms_ upload/media/1436/sict_artikel_engl.pdf)

DIN (2003b): Knappe Ressourcen effektiv nutzen. Strategieausschuss für die Standardisierung in der Informations- und Kommunikationstechnik (SICT) im DIN (http://www.sict.din.de/sixcms_upload/media/1437/sict_kb_strategieplan.pdf).

Friedlander, A. (2002): Dimensions and Use of the Scholarly Information Environment: Introduction to a Data Set Assembled by the Digital Library Federation and Outsell, Inc.. Digital Library Federation and Council on Library and Information Resources, Washington, D.C. (http://www.clir.org/pubs/abstract/pub110abst.html).

Harnad, S. (1991): 'Post-Gutenberg Galaxy: The Fourth Revolution in the Means of Production of Knowledge', ThePublic Access Computer Systems Review, vol. 1, no. 2, pp. 39-53.

Krause, J. (2006): 'Shell Model, Semantic Web and Web Information Retrieval', in I. Harms, H.-D. Luckhardt & H.W. Giessen (eds.): Information und Sprache. Beiträge zu Informationswissenschaft, Computerlinguistik, Bibliothekswesen und verwandten Fächern. Festschrift für Harald H. Zimmermann. K.G. Saur, München (in print).

Mann, Th. (2006): The Changing Nature of the Catalog and Its Integration with Other Discovery Tools. Final Report. March 17, 2006. Prepared for the Library of Congress by Karen Calhoun.

A Critical Review. April 3, 2006 (http://www.guild2910.org/AFSCMECalhounReviewREV.pdf).

Nardi, B. A., Zarmer, C. L. (1993): 'Beyond Models and Metaphors: Visual Formalisms in User Interface Design', Journal of Visual Languages and Computing, no. 4 (1993), pp. 5-33.

Nentwich, M. (2003): Cyberscience. Research in the Age of the Internet. Austrian Academy of Sciences Press, Vienna.

Poll, R. (2004): 'Nutzungsanalyse des Systems der überregionalen Literatur- und Informationsversorgung, Teil 1: Informationsverhalten und Informationsbedarf der Wissenschaft', Zeitschrift für Bibliothekswesen und Bibliographie, no. 51 (2004), pp. 59 - 75.

RSLG (2002a): Researchers' Use of Libraries and other Information Sources: Current Patterns and Future Trends. Final Report. Education for Change Ltd., SIRU, University of Brighton & The Research Partnership (http://www.rslg.ac.uk/research/libuse/).

RSLG (2002b): 'Research Support Libraries Group: Final Report', The New Review of Academic Librarianship, no. 8 (2002), pp. 3 - 86.

Stempfhuber, M. (2003): 'Adaptive and Context-Aware Information Environments based on ODIN – Using Semantic and Task Knowledge for User Interface Adaptation in Information Systems', in: Harris, Don; Duffy, Vincent; Smith, Michael ; Stephanidis, Constantine (eds.): Human-Computer Interaction: Cognitive, Social and ergonomic Aspects, Volume 3 of the Proceedings of HCI International 2003, 10th International Conference on Human Computer Interaction, 22-27 June 2003, Crete, Greece, pp. 864 – 868.

Stempfhuber, M., Hellweg, H., Schaefer, A. (2002): 'ELVIRA: User Friendly Retrieval of Heterogenous Data in Market Research', in: Callaos, N. at al. (eds.): SCI 2002: The 6th World Multiconference on Systemics, Cybernetics and Informatics; July 14-18, 2002, Orlando, Florida, USA, vol. 1, pp. 299-304

Stempfhuber, M., Kim, D-W., Petrick, M. (2003): 'Cross-Cultural Issues of Visual Formalisms in User Interface Design', in: Callaos, N. et al.(eds.): SCI 2003 - The 7th World Multiconference on Systemics, Cybernetics and Informatics, July 27 - 30, 2003, Orlando, Florida, USA, vol. 1, pp. 479 - 484.

Semantic Mapping of Chinese Geographical Classification Schemes

Xueying Zhang Guonian Lv Lin Qiu
(Jiangsu Provincial Key Lab of Geography Information Science, Nanjing Normal University, China)

Abstract

Semantic mapping of geographical classification schemes has been proved to be the most effective tool for solving the semantic inconsistency between heterogeneous geographical information systems. Intellectual mapping methods can normally achieve good performance, but problems always arise if there are insufficient time, money and domain experts. This paper proposes an automated method, which aims to effectively identify meaningful semantic relations of entities from different Chinese geographical classification schemes through a string similarity measure based semantic mapping model and a filtering algorithm.

Keywords

geographical classification scheme; semantic mapping; string similarity measure; geographical information system; semantic relation

Introduction

Geographical information has been widely applied in our daily life with the quick development of geographical information systems. A geographical classification scheme must be required for storage, search, management and analysis of geographical information in its system. Spatial location is the single

most important characteristic of geographical information; thus classifying geographical information has its difficulty and complexity. The types, attributes, relations and operations of geographical information must be taken into consideration as well as general classification rules (Liu & Jiang 2004). Since 80's four standard Chinese geographical classification schemes have been constructed, i.e. Classification and Codes for the National Land Information (GB/T13923-92)[1], Classification and Codes for the Features 1:500 1:1000 1:2000 Topographic Maps (GB/T14804-93)[2], Classification and Codes for the Features 1:5000 1:10000 1:25000 1:50000 1:100000 Topographic Maps (GB/T15660-1995)[3], and Classification and Codes for the Features 1:500 1:1000 1:2000 Topographic Maps for Urban Fundamental Geographical Information System (CJJ100-2004)[4]. Individual special Chinese geographical classification schemes are normally constructed for specific geographical information systems. GB/T13923-92 is a general and basic standard geographical classification scheme. All other Chinese geographical classification schemes are suggested to be compatible with GB/T13923-92.

Geographical information systems are usually inconsistent in terms of their content and geographical

[1] National Standard of the People's Republic of China. China State Bureau of Technical Supervision, 1992.
[2] National Standard of the People's Republic of China, China State Bureau of Technical Supervision, 1993.
[3] National Standard of the People's Republic of China. China State Bureau of Technical Supervision, 1995.
[4] Technical specification for urban fundamental geographical information system. Professional standard of the People's Republic of China. China Architecture & Building Press, 2004.

classification schemes. A single geographical concept might be represented in different ways so that geographical information systems cannot initially implement semantic information sharing. Semantic mapping is considered as a tool to state semantic relations between entities belonging to different classification schemes. It has been proven to most effectively resolve semantic heterogeneity between geographical classification schemes. Although intellectual methods can achieve good performance, problems always arise if there are insufficient time, money and domain experts.

There have been several proposals for automated drawing semantic mapping of thesauri, classifications and ontologies. Many of them define similarity measures based on parallel databases. For example, the log-likelihood ratio is used to produce LCSH heading/DDC number parings (Vizine-Goetz, 1998), and a rough set theory based model is proposed for integration of the concepts of document databases using various indexing languages (Zhang, 2006). String similarity measures are introduced to semantic mapping of ontologies (Ehrig & Euzenat, 2005; Maedche & Staab, 2002) . Recently a neural network and string similarity measures based data mining approach is used for recognition of relationships between elements of ontologies (Hariri et al., 2005). In this approach, one training set (i.e. one table) must be created for each pair of ontology, and each row shows comparison of an entity from first ontology to an entity from the second and existence of actual mapping between the two entities with true or false.

This paper presents an automated method for semantic mapping of Chinese geographical classification schemes without using parallel databases and training sets. The proposed method focuses on a string similarity based semantic mapping model and a filtering algorithm for effectively identifying meaningful relations of entities from different Chinese geographical classification schemes.

String similarity measures

In many applications, it is necessary to determine the similarity of two strings. The assumption is that there is a meaningful semantic relation between the two resembling strings. A widely used notion of string similarity is the Levenshtein distance (edit distance): the minimum number of insertions, deletions, and substitutions required to transform one string into the other (Levenshtein ,1966). (Afterwards) As time progress, improved measures have been proposed. For example, Needleman & Wunsch (1970) assigned a different cost on the edit operations; Smith & Waterman (1981) use an alphabet mapping to costs; Jaro (1995) presented a similarity measure which counts the common characters between two strings even if they are misplaced by a short distance. Monge & Elkan (1996) used variable costs depending on the substring gaps between the words. Stoilos et al.(2005) proposed a similarity trying to modify existing approaches for entities of an ontology. String similarity measures are sometime deceptive, when two strings resemble each other though there is no meaningful relation between them, such as “power” and “tower” (Maedche & Staab, 2002). The previous research shows that string

similarity measures are suitable for determining the similarity of two strings in English or other western languages.

There are great linguistic differences between Chinese and English. English sentence formations use spaces as its word delimiters while Chinese texts do not use spaces to mark word boundaries; a written Chinese sentence is a string of characters between punctuations. Since 90's attempts have been made to use appropriate string similarity measures for Chinese natural language information retrieval and automatic translation. Initially Wang & Wu (1993) proposed a location based string similarity measure (LSSM). Assuming that there are two strings S1 and S2, Match(S1, S2) is the number of common characters of S1 and S2, Num(S1) and Num(S2) are the character numbers of S1 and S2 respectively, a matching coefficient is denoted by α (default value 0.6), and a location coefficient is denoted by β (default value 0.4). The LSSM is defined by

$$Sim(S1,S2)=\alpha\times\frac{1}{2}\left(\frac{Match(S1,S2)}{Num(S1)}+\frac{Match(S1,S2)}{Num(S2)}\right)+\beta\times\min\left(\frac{Num(S1)}{Num(S2)},\frac{Num(S2)}{Num(S1)}\right) \quad (1)$$

In real-world applications, the definition of a threshold of Sim(S1, S2) is very difficult. Song (1996) proposed a SMSM measure for the definition of the similarity of two Chinese strings. Assumes that N denotes the number of common characters of strings S1 and S2, C1 denotes the proportion of N and the number of S1, and C2 denotes the proportion of N and the number of S2. The semantic similarity of S1 and S2 is defined as: if C1=1 and/or C2=1, then A and B are synonyms definitely; if 0.5≤C1<1 and/or 0.5≤C1<1, then A and B are synonyms roughly; if C1<0.5 and C2<0.5, then there is no meaningful semantic relation between S1 and S2. Although the SMSM is obviously

simpler than the LSSM, for a majority of strings the situation is similar to the second case, i.e. there still exists semantic uncertainty of S1 and S2. Wu (1999) tested the performances of the two measures in the same experimental condition. The result indicates that the LSSM is better than the SMSM. However, the LSSM does not take into account one significant Chinese characteristic of Chinese language, i.e. semantic center of gravity back. For most Chinese words, their second half characters are always more important than their first half characters in the representation of their semantic content. Thus Wu presented a GBSM measure defined by Formula (2).

$$Sim(S1,S2)=\alpha\times\frac{1}{2}\left(\frac{Match(S1,S2)}{Num(S1)}+\frac{Match(S1,S2)}{Num(S2)}\right)+\beta\times \\ \times\min\left(\frac{Num(S1)}{Num(S2)},\frac{Num(S2)}{Num(S1)}\right)\times\frac{1}{2}\left(\frac{\sum_{k=1}^{Match(S1,S2)}Location(S1,k)}{\sum_{i=1}^{m}i}+\frac{\sum_{k=1}^{Match(S1,S2)}Location(S2,k)}{\sum_{j=1}^{n}j}\right) \quad (2)$$

where m and n are the total number of S1 and S2 respectively, location(S1, k) and location(S2, k) are the location ordinal numbers of *k*th matching character in S1 and S2 respectively. For example, there are two strings "平面控制点"（horizontal control point, S1）and "一等平面控制点"（first order horizontal control point, S2）. Matching characters of S1 and S2 are "平"（flat）, "面"(surface), "控"(control), "制" (restrict), "点" (point). The semantic similarity of S1 and S2 is computed by

$$Sim(S1,S2)= \\ =0.6\times\frac{1}{2}\left(\frac{5}{5}+\frac{5}{7}\right)+0.4\times\min\left(\frac{5}{5},\frac{5}{7}\right)\times\frac{1}{2}\left(\frac{1+2+3+4+5}{1+2+3+4+5}+\frac{3+4+5+6+7}{1+2+3+4+5+6+7}\right)= \\ =0.78$$

（3）

Wu points out that the GBSM can achieve better performance than other measures for identifying semantic relations of Chinese domain terminologies. Class names of Chinese geographical classification schemes mainly use geographical terminologies to represent geographical information. Therefore, in this paper it is assumed that the semantic similarity of two entities from different Chinese geographical classification schemes could be effectively estimated in terms of the GBSM similarity of their class names.

Preprocessing of geographical classification schemes

There are multifarious flat and hierarchical facet classification schemes in information retrieval. In theory, hierarchical classification schemes can preferably represent qualitative geographical information; whereas facet classification schemes can better represent quantitative geographical information. Hierarchical facet classification schemes might be ideal for representation of geographical information. However, all existing Chinese geographical classification schemes are based on hierarchical architecture. This paper will discuss semantic mapping of Chinese hierarchical geographical classification schemes.

An entity of a geographical classification scheme mainly includes a class number and a class name. The class number is unique in a given scheme, i.e. the identification code of an entity. The class name is the literal description of semantic content of an entity. To help users easily remember and understand entities, class names are usually repetitive and compact. A class name can usually represent partial

semantic content of its corresponding entity. This is on the assumption that users can automatically understand an entity through analysis of its class name and context, especially its superclass name. The literal combination of a class name and its superclass name might more completely and correctly represent the semantic content of its corresponding entity. For this reason, class names of geographical classification schemes must be transformed as shown in table 1.

Furthermore, punctuations must be removed from the class names in the preprocessing of geographical classification schemes because they are uninformative in representation of semantic content of entities.

Class number	Class level	Class name	Transformed class name
1	First-level	测量控制点 (surveying control point)	测量控制点(surveying control point)
11000	second-level	平面控制点 (horizontal control point)	平面控制点测量控制点 (horizontal control point surveying control point)
11010	third-level	大地原点(geodetic origin)	大地原点 (geodetic origin surveying control point)
11020	third-level	三角点(triangulation point)	三角点 (triangulation point surveying control point)
11021	fourth-level	一等(first order)	一等三角点 (first order triangulation point)
11022	fourth-level	二等(second order)	二等三角点 (second order triangulation point)
11023	fourth-level	三等(third order)	三等三角点 (third order triangulation point)
11024	fourth-level	四等(fourth order)	四等三角点 (fourth order triangulation point)
11030	third-level	导线点(traverse point)	导线点 (traverse point surveying control point)
11031	fourth-level	一等(first order)	一等导线点 (first order traverse point)
11032	fourth-level	二等(second order)	二等导线点 (second order traverse point)

Table 1 Transformation of entities of a geographical classification scheme (GB/T13923-92)

Semantic mapping model of geographical classification schemes

Riesthuis (1996) argues that semantic mapping from classification scheme A (the source scheme) to classification scheme B (the target scheme) is not the same as semantic mapping from classification scheme B (the source scheme) to classification scheme A (the target scheme). Let E_A and E_B be the two entities from A and B respectively, the terms SE_A and SE_B denote class names of E_A and E_B respectively, and the terms TE_A and TE_B denote transformed class names of E_A and E_B respectively. The GBSM based semantic mapping model (GBSM-SM model) for semantic mapping of A (the source scheme) and B (the target scheme) is defined as follows:

M_1: $$Sim(E_A, E_B)_1 = Sim(TE_A, TE_B) \quad (4)$$

M_2: $$Sim(E_A, E_B)_2 = (1-\lambda)Sim(TE_A, SE_B) \quad (5)$$

M_3: $$Sim(E_A, E_B)_3 = (1-2\lambda)Sim(SE_A, TE_B) \quad (6)$$

M_4: $$Sim(E_A, E_B)_4 = (1-3\lambda)Sim(SE_A, SE_B) \quad (7)$$

M_1, M_2, M_3 and M_4 are in decreasing priority for the definition of $Sim(E_A,E_B)$ in terms of their correctness for identifying semantic relation of E_A and E_B. λ is a priority coefficient, which makes comparable M_1, M_2, M_3 and M_4.α is usually defined as 0.1.

Semantic mapping of two geographical classifications schemes A and B aims to identify meaningful semantic relations of their entities. However, for a given entity E_A of A, it will be correlated to every entity of B with four semantic similarity values in terms of the GBSM-SMM model. In this case, the problem is how to effectively filter an optimal mapping

entity of B for E_A. A filtering algorithm (see Table 2) is developed to solve this problem.

Input: E_{A1}, E_{A2}, …, E_{Ai} …E_{Am} are the entities of geographical classification scheme A, E_{B1}, E_{B2}, …, E_{Bj} …E_{Bn} are the entities of geographical classification scheme B.
Output: Semantic relations of A and B (Table T)
T=Ø
Define mapping coefficient α //Default value 0.6
Define location coefficient β // Default value 0.4
Define priority coefficient λ //Default value 0.1
Define similarity threshold θ //Default value 0.6
For i=1 to m
$E_A=E_{Ai}$
For j=1 to n
Temp table Q= Ø
$E_B=E_{Bj}$
Calculate $Sim(E_A, E_B)_1$ with M1
Add a semantic relation of E_A and E_B into table Q, $Sim(E_A,E_B)= Sim(E_A, E_B)_1$
j=j+1
Endfor
If max($Sim(E_A, E_B)$) in table Q >θ then
Select semantic relations from table Q into table T with max($Sim(E_A, E_B)$)
Endif
For j=1 to n
Temp table Q= Ø
$E_B=E_{Bj}$
Calculate $Sim(E_A, E_B)_1$ with M2
Add a semantic relation of E_A and E_B into table Q, $Sim(E_A,E_B)= Sim(E_A, E_B)_2$
j=j+1
Endfor
Count E_A from T into r where $E_A \neq Ø$
If r=0 then
Add records of Q with max($Sim(E_A,E_B)$) >θ into T
Else
Delete records of T with E_A and $Sim(E_A,E_B)$< max($Sim(E_A,E_B)$)

Add records of Q with max(Sim(E_A,E_B)) >θ into T
Endif
For j=1 to n
Temp table Q= Ø
E_B=E_{Bj}
Calculate Sim(E_A, E_B)$_1$ with M3
Add a correlated relation of E_A and E_B into table Q, Sim(E_A,E_B)= Sim(E_A, E_B)$_3$
j=j+1
Endfor
Count E_A from T into r where E_A≠Ø
If r=0 then
Add records of Q with max(Sim(E_A,E_B)) >θ into T
Else
Delete records of T with E_A and Sim(E_A,E_B)< max(Sim(E_A,E_B))
Add records of Q with max(Sim(E_A,E_B)) >θ into T
Endif
For j=1 to n
Temp table Q= Ø
E_B=E_{Bj}
Calculate Sim(E_A, E_B)$_1$ with M4
Add a correlated relation of E_A and E_B into table Q, Sim(E_A,E_B)= Sim(E_A, E_B)$_4$
j=j+1
Endfor
Count E_A from T into r where E_A≠Ø
If r=0 then
Add records of Q with max(Sim(E_A,E_B)) >θ into T
Else
Delete records of T with E_A and Sim(E_A,E_B)< max(Sim(E_A,E_B))
Add records of Q with max(Sim(E_A,E_B)) >θ into T
Endif
i=i+1
Endfor
Return table T

Table 2 A filtering algorithm based on the GBSM-SM model

Table 3 shows some mapping results of GB14804-1993 (the source scheme) and GB/T 13923-

92 (the target scheme) using an automated method based on the GBSM-SM model and its filtering algorithm. Each row represents one semantic relation of two entities.

E_A(GB14804-1993)		E_B(GB/T13923-92)		Sim(E_A,E_B)
Class number	Class name	Class number	Class name	
1	测量控制点(surveying control point)	1	测量控制点(surveying control point)	1.00
11	平面控制点(horizontal control point)	11000	平面控制点(horizontal control point)	1.00
111	三角点(triangulation point)	11020	三角点(triangulation point)	1.00
1111	一等(first order)	11031	一等(first order)	1.00
1112	二等(second order)	11032	二等(second order)	1.00
1113	三等(third order)	11033	三等(third order)	1.00
1114	四等(fourth order)	11034	四等(fourth order)	1.00
112	土堆上的三角点 (triangulation point on hillock)	11020	三角点(triangulation point)	0.74
1121	一等(first order)	11031	一等(first order)	0.75
1122	二等(second order)	11032	二等(second order)	0.75
1123	三等(third order)	11033	三等(third order)	0.75
1124	四等(fourth order	11034	四等(fourth order	0.75
113	小三角点(minor triangulation point)	11020	三角点(triangulation point)	0.91
114	土堆上的小三角点 (minor triangulation point on hillock)	11020	三角点(triangulation point)	0.71
115	导线点(traverse point)	11030	导线点(traverse point)	1.00

1151	一级(first class)	11031	一等(first order)	0.75
1152	二级(second class)	11032	二等(second order)	0.75
1153	三级(third class)	11033	三等(third order)	0.75
116	土堆上的导线点(traverse point on hillock)	11030	导线点(traverse point)	0.71
117	埋石图根点 (monumented mapping control point)	13052	埋石图根点 (monumented mapping control point)	0.87
118	不埋石图根点 (unmonumented mapping control point)	13052	埋石图根点 (monumented mapping control point)	0.62
12	高程控制点(vertical control point)	12000	高程控制点(vertical control point)	1.00
121	水准点(benchmark)	12020	水准点(benchmark)	1.00
1211	一等(first order)	12021	一等(first order)	1.00
1212	二等(second order)	12022	二等(second order)	1.00
1213	三等(third order)	12023	三等(third order)	1.00
1214	四等(fourth order)	12024	四等(fourth order)	1.00
1215	五等(fifth order)	12024	四等(fifth order)	0.77
13	GPS点(GPS point)	13030	GPS点(GPS point)	1.00
14	其他控制点(other control point)	13000	其他控制点(other control point)	1.00

Table 3 Results of semantic mapping of two geographical classification schemes

Experimental evaluation

The conclusion of semantic relevance between entities of classification schemes is subjective in any real-world task. Nevertheless, there is no doubt that intellectual methods can achieve the best performance. Four parameters will be introduced to estimate the performances of intellectual and automated methods for semantic mapping of geographical classification schemes.

Recall (R) and precision (P) are the basic parameters used in evaluating search strategies (Salton, 1983). R is the ratio of the number of relevant records retrieved to the total number of relevant records in the database. P is the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved. In this paper R and P are redefined by

$$\text{Recall} = \frac{\text{The number of unique source entities of the semantic relations created by one method}}{\text{The total number of entities of the source classification scheme}} \quad (8)$$

$$\text{Precision} = \frac{\text{The number of unique target entities of the semantic relations created by one method}}{\text{The total number of entities of the target classification scheme}} \quad (9)$$

Here R and P address the compatibility of two geographical classification schemes through semantic mapping implementation. Accuracy ratio (AR) is the proportion of correct semantic relations out of the total semantic relations created by one automated method. All automatic results are reviewed by domain experts. Matching ratio (MR) expresses the proportion of overlaps of semantic relations generated by the automated method from the whole semantic relations created by the

intellectual method. The matching ratio measures the ability of one method to replace the intellectual method.

Our experiments are carried out on four Chinese geographical classification schemes, i.e. GB/T 13923-92 (B1), GB 14804-1993 (B2), GB/T 15660-1995 (B3) and a specific scheme (NT in short). The semantic mapping of NT and GB/T 13923-92, NT and GB 14804-1993, and NT and GB/T 15660-1995 have been implemented intellectually, where the NT is the source scheme and the B1, B2, and B3 are the target schemes. The experimental estimation results are shown in Table 4. It should be pointed out that all coefficients in the GBSM-SMM model and the filtering algorithm are defined as default values in our experiments.

Initially, it can be seen from Table 4 that the automated method achieves a little worse recall and a little better precision than the intellectual method on the three experiments. However, the number of semantic relations created by the automated method is lower than the number of semantic relations created by the intellectual method. The main reason is that our mapping model is based on string similarity measure, and some transformed class names can still represent partial semantic content of their corresponding entities. Moreover, string similarity measures always ignore the indeed semantic relations of geographical terminologies, especially these semantically correlated entities but with lower string similarity.

Secondly, a satisfactory accuracy ratio is available for the automated method. In particular, the accuracy ratio of automatic mapping results of NT and GB14804-1993 is greater than 96%, which signifies that the automatic results can be corrected with little human interruption

classification scheme		Number of entities		Number of semantic relations		Unique entities (Intellectual Method)		Unique entities (Automated method)		Intellectual Method (%)		Automated method (%)		AR (%)	MR (%)
Source	Target	Source	Target	IM	AM	Source	Target	Source	Target	R	P	R	P		
NT	B1	1058	661	829	620	809	588	615	580	76.5	88.9	58.1	87.7	84.4	73.5
NT	B2	1058	663	866	885	623	628	634	630	58.9	94.7	59.9	95.0	96.7	33.2
NT	B3	1058	603	761	596	725	545	592	555	68.5	90.4	56.0	92.0	84.4	61.7
Average		1058	642	818	700	719	587	614	588	68.0	91.3	58.0	91.6	88.5	56.1

Table 4 Experimental evaluation results

In addition, we can also seen from Table 4 that the matching ratio varies from 33.2% to 74.%. On the experiment of NT and GB14804-1993, the automated method finds many semantic relations that are not addressed by the intellectual method. It indicates that the combination of two methods could achieve better semantic mapping performance than a solo method.

Conclusion and future work

This research has shown that the GBSM-SM model and its filtering algorithm can effectively and automatically link the semantic relations of entities from different Chinese

geographical classification schemes. Semantic mapping is significant for modification and integration of geographical information systems.

Our future research would focus on a knowledge base of geographical terminologies and semantic mapping of Chinese geographical ontologies using our proposed automated method.

Reference

Doerr, M. (2001). Semantic problems of thesaurus mapping. Journal of Digital Information, 1(8), No.52.

Ehrig, M. & Euzenat, J. (). Relaxed precision and recall for ontology matching. In *Proceedings of K-Cap 2005 Workshop on Integrating Ontology* (pp. 25-32).

Maedche, A. & Staab, S. Measuring similarity between Ontologies. In *Proceedings of the European Conference on Knowledge Acquisition and Management* (pp. 251-263).

Liu, R. & Jiang, J. (2004). Classification principles and methods of geographical information-A case study on basic geographical information. Science of Surveying and Mapping, 29(7), 84-87.

Needleman, S.B. & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. Molecular Biology, 48, 443-453.

Zhang, X. (2006). Concept integration of document databases using different indexing languages. Information Processing & Management, 42, 121-135.

Stoilos, G. et al. (2005). A string metric for ontology alignment. In *Proceedings of the 9th IEEE International Symposium on Wearable Computers* (pp. 224-237)

Jaro, M. (1995). Probabilistic linkage of large public health data files. Molecular Biology, 14, 491-498.

Riesthuis, G. J. A. (1996). Theory of compatibility of information languages. In *Compatibility and Integration of Order System,Research Seminar Proceedings of the TIP/ISKO Meeting* (pp. 23–31).

Monge, A. E. & Elkan, C.P. (1996). The field-matching problem: Algorithm and applications. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining* (pp.267-270).

Smith, T. F. & Waterman. (1981). Identification of common molecular subsequences. Molecular Biology, 147, 195-197.

Riesthuis, G. J. A. (1996). Theory of compatibility of information languages. In *Compatibility and Integration of Order System, Research Seminar Proceedings of the TIP/ISKO Meeting* (pp. 23–31).

Salton, G., & McGill, M. J. (1983). Introduction to modern information retrieval. New York: McGraw-Hill Book Company.

Senellart, P. (2001). Extraction of information in large graphs. Automatic search for synonyms. Technical report, Universite catholique de Louvain, Belgium.

Resnik, P.(1995). Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence* (pp. 448-453).

Levenshtein, I. V. (1966). Binary codes capable of correcting deletions, insertions, and reversals. Cybernetics and Control Theory, 10, 707-710

Vizine-Goetz, D. (1998). Subject headings for everyone: Popular Library of Congress Subject Headings with Dewey Numbers. OCLC Newsletter, 233, 29–33.

Wang, Y. & Wu, X. (1993). Automatic processing of post-controlled rules. New Technology of Library and Information Service, (2).

Wu, Z. (1999). Study on economic information post-controlled thesaurus. Master's degree, Nanjing Agricultural University.

Song, M. (1996). Study on Chinese word literal similarity and post-controlled thesaurus. Journal of the China Society for Scientific and Technical Information, (4), 261-267.

Hariri, B. B. et al. (2006). Combining ontology alignment metrics using the data mining techniques. In *Proceedings of International Workshop on Context and Ontolgies.*

Exploring the Cloud of Research Information Systematically

Keith G Jeffery
CCLRC;
keith.g.jeffery@rl.ac.uk

Anne Asserson
UiB
anne.asserson@fa.uib.no

Abstract

Imagine a cloud of data points (values) in multidimensional space. This is how research information appears commonly to end-users. There are many different kinds of data with varying quality, structured, semi-structured or unstructured with heterogeneous character sets, languages, syntax and semantics.

Different end-users see the cloud differently: some data points are of more interest, some of less interest. The 'more-or-less' dimension has several aspects: the kind of data (entities), the values of data points (attribute-values) and the relationship between sets of data values (such as relationship role or relationship duration).

As one example, a typical researcher is extremely interested in a scientific dataset and associated software together with the associated metadata describing the precision, accuracy, method of collection etc. The researcher is very interested in associated publications with interpretations of the scientific dataset. The researcher will be interested in who conducted the research, the rest of the team and the organisation where

the research was done. The researcher has less interest in who funded the research, under which programme it was funded. From this we can construct a set of 'axes' or structural elements through the cloud along which data points of interest cluster.

As another example, a research manager in a funding organisation may be interested in the value of funding awarded within one country to research in a particular subject area without being interested in to whom it was awarded nor the organisation where they work. This is a different view of the cloud with a different set of axes along which data points cluster. A research manager at a university might also be interested in output research publications by year, by department, by publication channel.

As a final example consider the innovative entrepreneur. She will be interested in any products from or patents on research in a relevant topic across all countries, on the track-record of a researcher or her organisation in technology transfer and wealth creation and in any conditions attached to the research funding. This is yet another set of axes through the cloud.

Three questions are paramount: (1) how to assure quality data (so that results are accurate); (2) how to assist the end-user in formulating correctly the query to obtain the expected results; (3) how to structure the data to obtain the optimal response in terms of performance, recall and relevance including taming heterogeneity.

We demonstrate that all three are related. We outline a solution based on structured metadata, formal logic and knowledge engineering techniques exposed to the end-user as a user-friendly assistant with graphical metaphors.

The Problem

Research Information covers a broad field. It includes unstructured or semistructured information such as national and funding organisation strategic papers, programmes and funding and the full text of research output publications, It includes structured information such as that used for management of research in funding organisations and research institutions, and also research datasets. In different countries the information may be encoded with different character sets and in different languages. The terminology – used as entity or attribute names for example, or as valid values of an attribute - may be different from country to country and even organisation to organisation.

The end-user wants a homogeneous response to a query (which may involve functional processing in addition to a simple retrieval). She wants the response in a reasonable time and with all relevant information (recall) and perhaps some indication of relevance (how closely he answer matches the query).

The problem is actually based on three key parameters: data quality, query precision and data structure.

Data Quality

It is a fact of modern life that the real world is represented in computer systems and management decisions are based upon the information in those systems. Air traffic controllers do not look out of the window at planes but see a representation on radar screens. Financial analysts do not see the product stock in the warehouse, the piles of raw material or the actual

production line – just the data. The same is true of users of CRISs.

The information used for decision-making is data structured in context. The data quality is paramount: poor data (i.e. data that represents inaccurately the real world) leads to poor decisions (which are effected by actions in the real world). The more processed the data (e.g. through retrieval, analysis, models and simulation with graphical presentation) the more inaccuracies in the data are magnified. Mercifully air traffic control data is more-or less a simple radio message represented on screen.

User Query

The end-user commonly has a query conceptualised in her own mind-space using her own terminology, or (better) that commonly used across the subject domain. The requirement may be to compare the research performance of their university against others across a range of metrics (products, patents, publications) over years. How does the user express this to a CRIS? For sure she is not going to sit down and write the SQL query and – since SQL is relationally complete but not functionally complete – she would have to write or use pre-existing software to do the analytical part of the query including any graphical display. How should she express the query such that the output compares like with like: are the years calendar or academic? Are the patents measured by number or licence value?

Data Structure

The end-user requires a timely answer. She requires relevant information to her query. She needs to know if she has all relevant information (recall) or only

some of it – and the degree of relevance. All these properties of information are influenced heavily by data structure. Structured information can be ordered, indexed (multiply), clustered physically or partitioned as structured streams. If data are not structured, it is not possible (within the closed world of the CRIS) to determine recall. If data are not structured, it is not possible to determine relevance. If data are not structured, it is difficult to optimise queries. Despite claims concerning semi-structured data and information retrieval (where the concepts of recall and relevance originated) in fact the calculations to determine these characteristics are performed on structured metadata, not on the semi-structured or unstructured data.

Problem Conclusion

The different kinds of user requirement outlined in the abstract all require answers to the three questions posed there and elaborated in this section. We now consider a sample of related work (the general literature is copious, that related to CRIS moderate) and then offer and analysis and solution before concluding with a roadmap of future work.

Related Work

There is a plethora of related work on structured data, databases and information retrieval, relating to effectiveness and efficiency of representation of the real world. From basic textbooks e.g. [Da04] to detailed technical papers e.g. the VLDB Conference series [VLDB] there is wide discussion of these issues. Similarly, query optimisation is covered in detail in e.g. the VLDB Conference series [VLDB]. User interface

issues are also covered widely, perhaps most comprehensively in the CHI Conference series [CHI].

More important for CRIS is the application of this background – and generally theoretical - knowledge to the CRIS domain. The CRIS Conference series contain a wealth of relevant papers including those by the authors. There have been two major approaches:

integrating heterogeneous distributed databases of CRIS information (including associated open access institutional repositories and research datasets – the latter represented by structured metadata) into a homogeneous canonical form;

harvesting semistructured information from the web to create a structured metadata index (with limited information in a structured form) which points to the original semistructured data in its native form;

Clearly (a) provides the end-user with information in a consistent form that can be used for comparison and calculation whereas (b) does not. However, (a) requires a different and more disciplined approach upstream to input – best managed in a workflow environment [JeAs06a] – and consequent less effort downstream in retrieval, integration and analysis. In contast, (b) takes any input (usually semistructured or unstructured) with low cost upstream and attempts (by human effort and therefore high cost) to make sense of it downstream.

The conclusions we draw from this related work are:

1. unstructured or semistructured data is valuable information but is only of real value – in terms of quality, accessibility and performance - when indexed by structured metadata [Je98] and then processed using appropriate specialised techniques to provide information and extract knowledge;

2. satisfactory responses to end-user queries (relevance, recall, performance) can only be produced when the data is structured, or other kinds of data are indexed by structured metadata;

3. satisfactory response to end-user queries over heterogeneous distributed data (the usual case for any real-world query) requires knowledge-based techniques for schema description, reconciliation and explanation [JeHuKaWiBeMa94] ;

The Solution

The solution relies on technologies that address the topics raised in the problem statement. All three require quality and structured information. All three require knowledge engineering techniques.

Quality Structured Information: CERIF

Data

Data quality can only be assured if the input or edit of data attribute values is validated and – if necessary – supported with explanation of valid values. To achieve this requires structured data i.e. data arranged as attribute values in a structure. There are many validation techniques and many are applicable to each data attribute [GoGlJe93]. Most validation relies on first order logic and Boolean algebra. This demands structured data.

CERIF [CERIF] provides a data model constructed and maintained by international experts in the requirements of CRIS. It is continually being improved and extended by a well-defined process. It has been used in any CRISs and demonstrated to be effective and efficient. It covers the major entities and attributes of a CRIS. Recent standardised extensions handle more

detailed bibliographic information, while experts are experimenting with extensions to provide more financial information, more information related to innovation and information to manage e-processes.

Metadata

The highly structured and optimised CERIF model can act as metadata to other structured, semistructured and unstructured information as demonstrated in [JeAs06]. Examples include Open Access Institutional Repositories of publications [AsJe05] and research datasets and software. This ensures that such datasets are validated, retrieved and interpreted in a structured and logical context.

Knowledge Engineering

Expert Advisor and Query Assistant

The end user commonly has difficulty in formulating a query appropriately, particularly if the query is complex involving functions (from simple COUNT, SUM, AVG through to user-defined complex statistical analyses or modelling). Here an expert advisor can help, by deprecating inappropriate function use (e.g. a numerical function over attributes whose values are of type character), by assisting with suggested functions available, by assisting the query optimiser through screening first inappropriate use of the relational calculus, by reminding the end-user of available entites and attributes, by providing graphical comparisons of schemas to assist in semantic reconciliations etc. The expert advisor can only function effectively and efficiently on structured data where the data quality is good.

Expert advisor systems rely on several components: a human-computer interaction component

which may be multilingual and multimodal; a domain ontology containing knowledge of the domain of interest; an inference engine to do the required logical processing and an interface to the underlying processing systems such as database query, statistical analysis, graphical visualisation etc. With these components the systems react intelligently (and sometimes annoyingly) to the end-user with the intention of causing the underlying processing system to meet more exactly the user requirement. The underlying processing systems may handle a mix of structured, semistructured and unstructured data but the optimisation of the user request requires structured, logical data.

Homogeneous View: Knowledge-based Information Integration

The end user usually requires a clear, structured answer: a homogeneous view over heterogeneous data. Starting from heterogeneous distributed databases is not ideal: there are so many possible differences of character sets, language, naming, typing, domain constraints, structure etc. The provision of the homogeneous view has been an open research question for > 30 years and has attracted prodigious research efforts. Indeed, the lack of an elegant solution to this problem led, in a way, to the use of WWW, harvesting and the end-user to browse and select intelligently. However, this approach is very expensive (person time), error prone (ambiguities in character set, language, syntax and semantics in web resources) and inefficient in use of computing resources.

Reconciliation of heterogeneity requires knowledge assisted arbitrage: the end user is presented graphically with the different schemas with attribute names (for example) and the system proposes likely equivalences which the user – using a graphical editor -

can correct manually which leads to learning by the system. The knowledge-based system relies on structured data to operate.

One demonstrated technique from the healthcare domain involves matching schemas of native databases to a common canonical form using a domain ontology and inference engine, and then generating software to convert from the native form to the canonical form [SkKoBeJe99]. The advantage of using the canonical form is, of course, that it reduces a n*m problem to a n problem. For the CRIS environment a suitable canonical form is CERIF [Je05].

Explanation

Once an end-user has received an answer from a CRIS system it is likely that there is a requirement for interpretation of the answer. It is here that an expert advisor system can assist, by explaining how the system derived the answer from the base information and the required processing [WiChLa93]. This is usually an improvement on the end-user guessing the explanation and is more effective (because of the use of domain knowledge and therefore context) and efficient than the end user searching the web for an explanation.

Analysis, Modelling, Visualisation

It is usually insufficient for the end-user to receive the output from a database query. Commonly the end-user wishes to understand the significance of the result which will involve further processing such as statistical analysis or modelling for prediction. The volumes and complexity of the information so produced can be daunting; it is here that visualisation (graphical representation) becomes valuable to provide the end-user with views on and insights into the information

hitherto obscured. Graphs showing attribute value set characteristics (maxim, minimum) and inter-relationships (eg correlation coefficient best-fit line) are valuable, as are multivariable diagrams and 'movies' showing changes with time or in any other dimension.

Push Technology

The above all is predicated upon the end-user initiating the request to the system. However, push technology allows a user profile to be stored and for relevant information, in an appropriate form, to be supplied to the end-user whenever the state of the information system changes such that something of interest to the end-user appears. This is a very powerful use of technology since it means that the end-user is alerted to anything interesting and does not have to initiate the processing herself. The user profile and the generation of appropriate information to be 'pushed' depends on structured data, knowledge-based system technology and requires a modern distributed and ambient environment for communication and information transmission.

Utilisation of the Solution

Let us consider how the solution outlined above could be used by our three kinds of users introduced in the abstract.

The researcher would find an easy-to-use assisted interface; behind which integration of information from multiple sources would be performed automatically, even bringing into a homogeneous context semistructured information (notably publications and research datasets with associated software) via the structured metadata.

Thus the cloud of research information becomes a well-formed set of quality information.

The research manager in a funding organisation would also encounter an easy-to-use assisted interface with the required information from multiple heterogeneous sources integrated. The interface provides appropriate functions for comparing the funding and the explanation engine provides background information on the way in which funding is calculated in each country. This provides the manager with quality information structured in context and with appropriate explanation to assist in interpretation. The university research manager is assisted by te interface to formulate the query to ensure the correct year(s) and departments are selected and the count of publications is done correctly according to the criteria (e.g. against a list of peer-reviewed publication channels).

The innovative entrepreneur utilises the advanced interface to formulate a homogeneous query over the heterogeneous national sources of information. The query is improved by the inference engine and domain ontology to overcome the fact that terminology differs from country to country and she wishes to compare like with like in terms of patents, products and their generated wealth through licences or sales. More complex is to compare track records in wealth creation of research groups; a graphical representation of value against time is used with annotation to explain the basis of the reasoning leading to the inclusion or exclusion of certain research outputs.

Thus we can see that the requirements can be met by utilising a structured data model (CERIF) and knowledge-based information systems engineering for

query improvement, heterogeneous information integration and explanation.

Future Work

We have demonstrated across the requirements of data quality, query improvement and integration of information the common solution in structured information (as data or metadata) and knowledge-based techniques. The key technological challenge remains the homogeneous view of heterogeneous information. The attempts to solve the problem using the end-user as the knowledge-based component of the system are shown not to scale, do not allow easy integration of information retrieval and processing and also are error-prone. We suggest the solution to this problem utilising CERIF as the canonical data model with knowledge-based techniques for schema-matching is the highest priority and can deliver the greatest rewards for CRIS.

References

[CHI] http://www.chi2006.org/index.php

[Da04] C.J.Date: An Introduction to Database Systems ; Addison-Wesley 2004 ISBN-13: 9780321197849

[AsJe05] A Asserson, K G Jeffery: 'Research Output Publications and CRIS' The Grey Journal volume 1 number 1: Spring 2005 TextRelease/Greynet ISSN 1574-1796 pp5-8

[GoGlJe93] C A Goble, A Glowinski, K G Jeffery: 'Semantic Constraints in a Medical Information System' Proceedings BNCOD-11 'Advances in Databases' July 1993 pp40-57 Edited by Worboys,M and Grundy,A F; Lecture Notes in Computer Science Series 696, Springer Verlag, 1993

[JeHuKaWiBeMa94] K G Jeffery, E Hutchinson, J Kalmus, M D Wilson, W Behrendt, C A Macnee;: 'A Model for Heterogeneous Distributed Databases' Proceedings BNCOD12 July 1994; LNCS 826 pp 221-234 Springer-Verlag 1994[Je98] K G Jeffery: 'Metadata' Invited Paper CRIS98 Conference, March 1998, Luxembourg.

[Je99] Jeffery, K G: 'An Architecture for Grey Literature in a R&D Context' Proceedings GL'99 (Grey Literature) Conference Washington DC October 1999

[JeAsReKo00] K G Jeffery, A Asserson, J Revheim, J Konupek: 'CRIS, Grey Literature and the Knowledge Society' Proceedings CRIS2000 Conference, Helsinki, Finland May 2000 http://www.cordis.lu/cris/cris2000

[Je05] K G Jeffery CRISs, Architectures and CERIF CCLRC-RAL Technical Report RAL-TR-2005-003 (2005)

[JeAs06] Keith G Jeffery, Anne Asserson: 'CRIS Central Relating Information System' in Anne Gams Steine Asserson, Eduard J Simons (Eds) 'Enabling Interaction and Quality: Beyond he Hanseatic League'; Proceedings 8th International Conference on Current Research Information Systems CRIS2006 Conference, Bergen, May 2006 pp109-120 Leuven University Press ISBN 978 90 5867 536 1

[JeAs06a] Keith G Jeffery, Anne Asserson: 'Supporting the Research Process with a CRIS' in Anne Gams Steine Asserson, Eduard J Simons (Eds) 'Enabling Interaction and Quality: Beyond the Hanseatic League'; Proceedings 8th International Conference on Current Research Information Systems CRIS2006 Conference, Bergen, May 2006 pp 121-130 Leuven University Press ISBN 978 90 5867 536 1

[SkKoBeJe99] Skoupy,K; Kohoutkova,J; Benesovsky,M; Jeffery,K G: 'Hypermedata Approach: A Way to Systems Integration' Proceedings Third East European Conference, ADBIS'99, Maribor, Slovenia, September 13-16, 1999, Published: Institute of Informatics, Faculty of Electrical Engineering and Computer Science, Smetanova 17, IS-2000 Maribor, Slovenia,1999, ISBN 86-435-0285-5, pp 9-15

[WiChLa93] G Wickler, H Chappel, S C Lambert : An architecture for a generic explanation component Proc. IJCAI Workshop on Explanation and Problem Solving, Chambery, August 1993, p.53-64, M R Wick (ed.)

[VLDB] http://www.informatik.uni-trier.de/~ley/db/conf/vldb/index.html

Full Text Retrieval Systems, XML and Databases Where are Future Information Architectures heading?

Stefan Baerisch
IZ Sozialwissenschaften, Deutschland
bs@bonn.iz-soz.de

Abstract

Today's information architectures often have to provide support for highly structured data as well as nonstructured texts it is thus important that such systems are capable of traditional data retrieval tasks as well as of information retrieval. While the past saw considerable differences in both the architectures and the use cases for full text retrieval systems and databases, recent developments have narrowed this gap considerably. Current Retrieval Systems like the FAST ESP system support fieldbased search, numeric data and XML retrieval as well as a query language are sufficient for most use cases. Native or hybrid XMLDatabases provide fast access to structured, semi structured and nonstructured information using the XQuery Language while relational database vendors are adding XML Support to their products. This paper gives an overview of the potentials and limitations of the relevant technologies as well as an outlook of future development like the XQuery fulltext extension, which aims to bring ranking and thesaurus support to XMLDatabases. Scenarios for the integration of the diferent data models and technologies are given, including best practices for adding support for semistructured and non-structured data to existing information architectures.

Introduction

In the past, data retrieval and information retrieval were mostly separate domains, either concerned with the retrieval of data from highly structured databases using

exact query languages or with the interpretation and fulfilment vaguely formulated informations need on unstructured text. Recent years saw the narrowing of the gap between this domains, caused in no small part by the growing importance of the World Wide Web in nearly all aspects of life. Two aspects are of special significance: The rise of XML and the concept of relevance.

- XML differs from the relational data model defined by Codd [Cod70] as well as from unstructured text in that it is semi structured, thus making it possible for a schema to only loosely describe the data.
- The concept of relevance in turn is essential in order to work with large amounts of information. It is no longer sufficient for an information system or database to return correct results, it is also necessary to return important results first, otherwise the user of an information system will be presented with too many results without an indication which are the ones important to him.

The subject of this paper are trends in software systems developed for the handling of semi structured data as well as for the fulfilment of vague information needs. Different types of systems, such as relational databases, XML data management systems and full text retrieval systems are introduced, their properties and possible ways for integrating the systems into a common architecture are discussed.

This text is structured as follows: Section 2 explains the differences between data retrieval and information retrieval and gives a definition of semi structured data. Relational databases, XML data management systems and text retrieval systems are

introduced, following the definition of the retrieval tasks and the different data models. Section 3 discusses use cases which the different systems may encounter, with the focus on data integration. The strengths and weaknesses of the systems are given. Section4 discusses how the different systems and data models can be integrated in an unified information system. Section 5 concludes.

This report is based on the experiences the Gesis Social Science Information Centre gained in the development of the SOWIPORT portal. In the course of the preparation for this project an extensive evaluation of di□erent XMLdatabases and full text retrieval architectures was done.

Data models and Architectures

A data model defines the kind of data that can be stored in a given system as well as the operations possible on the data. While some systems o□er support for multiple data model, in most cases there will be only one native data models used to internal storage. This section introduces the different data models and the systems implementing them.

Data-Retrieval versus Information Retrieval

More and more databases are searchable on the web. The term deep web or hidden web [RGM01] was coined for resources which, while available for search by user, are not indexed by the common search engines. With this large number of databases online, it becomes increasingly important for each database to be able to assess and fulfill the users information need it can no longer be assumed that users are familiar with the database structure and query language. As a

consequence databases must offer not only capabilities for data retrieval but also for information retrieval.

Structured, Unstructured and semi structured Data

Semi structured data [ABS00] lies between fully structured data as found in relational databases and unstructured text in terms of structural information. It differs from unstructured data in that it may have a schema, either implicitly or explicitly. Other than with structured data however, this schema is not strict it may only describe the general structure of the data. A common example for semi structured data are web pages. While most web pages conform to the HTML schema, the schema itself imposes few restrictions on the content of the pages.

XML can be seen as the currently most frequently used form of semi structured data. A number of di□erent schema languages are available to specify schemas, the most widely used being XML Schema [vdV02] with DTDs and Relax NG [vdV03] as alternatives. Since the choice of the right schema language is important when working with XML, these three languages will receive a short description.

DTD

The Document Type Definition was the first Schema Language available for XML, it was developed from similar technology for SGML. DTD's main advantages are its simplicity and its wide support. It's main shortcomings are the lacking support for XML namespaces and data types, for instance no numeric values are supported. The fact that DTDs can not be expressed in XML themselves is also sometimes cited as a problem.

XML Schema

XML Schema was developed in part as a reaction to the lacking features in DTD. The language supports a powerful type system and supports inheritance. XML Schemas are XML documents, they therefore benefit from the larger number of XMLaware tools. An argument sometimes voiced against XMLSchema is the complexity and verbosity of the language. The support for a large number of features means that XML Schema itself takes some time to learn, while the XML syntax results in longer document than it would be the case with DTDs. Tool support for XML Schema is very good, most XMLParser include solid XML Schema support and a large number of editors are available for the structured editing of schemas.

Relax NG

Relax NG is the least well known of the schema languages introduced in this paper. While Relax NG is used for relevant project, for example it is used for the XHTML 2.0 [IDB+06] standard, it does not have the tool support of DTD or XML Schema. What makes Relax NG interesting is the fact that it is nearly as powerful as XMLSchema[1], but retains a relatively simple syntax with can both be written as text or XML.

Relational Databases

The underlying data model of relational databases as described by Codd [Cod70] can be simplified as a set of tables or relations. Complex data structures are expressed by references between tables in the form of

[1] 1Relax NG is a nondeterministic Schema Language [vdV03] allowing it to express some concepts not possible in XML Schema

primary keys. Support for XML and information retrieval is added to the data model by the way of extensions, often specific for a single vendor. The extensions add new elements to the SQL language and provide new data types. SQL2003 attempts to standardize a common set of data fields and query language extensions. Most current databases added support for XMLFields in the last years. For the implementation of the XML data model on a relational database engine, two approaches were commonly used:

Use of CLOBS XML content is regarded as a string and parsed on demand, Queries may be supported by special indices.

Shredding XML is parsed into a tree where the different elements are stored as separate entries in the database. Figure 1 shows shredding into a sidetable as done by the IBM DB2 database

Since both approaches can not guarantee sufficient performance, especially for large documents, most databases include special indices for XML. Some databases, called hybrid database like the IBM DB2 Viper, support both XML as well the relation data model through special storage structures.

As with XML, support for relevance ranking of the results for a certain query as well as for the ranking of results in general is provided by vendor specific extension.

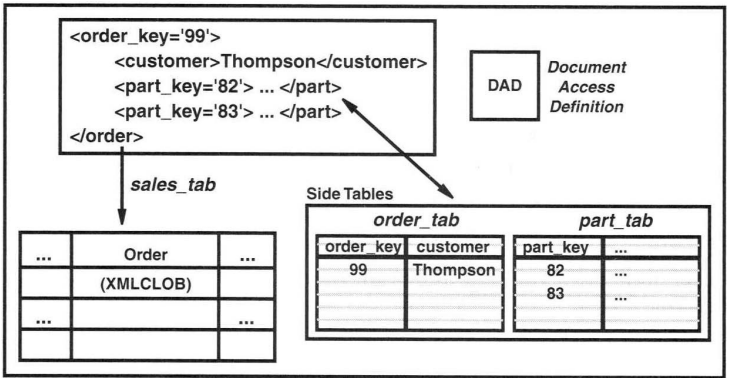

Figure 1: Shredding XML Content [Ben03]

XMLDatabases

XML databases or data management system support XML as their primary data model. Sometimes a distinction is made between native and nonnative XML databases, where the former use a specialized internal data structure to store XML while the latter depend on an internal data model other than XML to store documents. This distinction is somewhat artificial, since one of the most important characteristics of data management systems is the abstraction of the internal data representation. The eXist XMLDatabase for example supported different storage implementation in an earlier version as shown in figure 2. The API was not affected by the differences in the storage engines. The use of relational database for storing XML and the best representation of the data is a widely researched topic, see for example [LVLG03, ABS00, KKN03].

Given the definition of a XMLDatabase as a data management system with XML as the primary data model, the various XML database implementations exhibit large difference in their implementation of features [Sne05]. One of the most important characteristics of a

XMLdatabase is the Query language supported. This language determines not only the available operations for retrieval and the manipulation of data, but also has great influence on the data model [ST01]. If vendor specific and obsolete languages are not taken into account, two different query languages are of importance, XPath and XQuery.

The XPath Language

While the XPath Standard has reached Version 2.0 [FSC+06], many implementations are still based on the XPath 1.0 [CD99] standard. The XPath 1.0 language and type system is relatively simple while XPath 2.0 is a true subset of XQuery, including the complete XQuery type system.

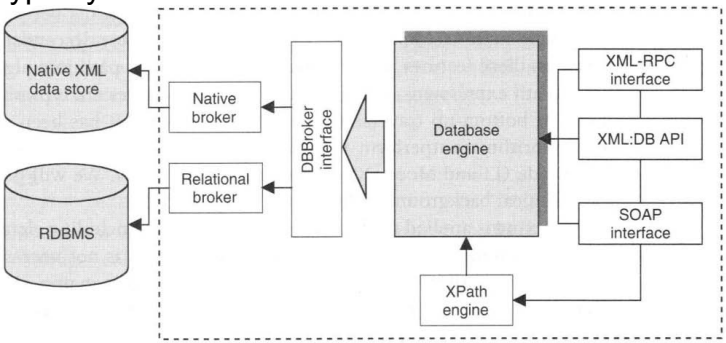

Figure 2: Storage Option for a version of the eXist database [Mei03]

Both versions of XPath are based on path expressions. Nodes to be retrieved from a XML document are specified by the names and values of other nodes in relation to them. The expression /Document/Author/Firstname for example will start at the current documents root and return all elements with the name Firstname which are found under the element Author which are in turn found under the element

Document. Path expressions can be restricted with predicates, the expression //Author[@id='1']/Firstname for instance returns only Author elements which are children of Author elements with an id attribute with the value 1. Figure 3 shows how elements are selected from a XML document by the XPath expression /customers/customer/invoice/item/quantity.

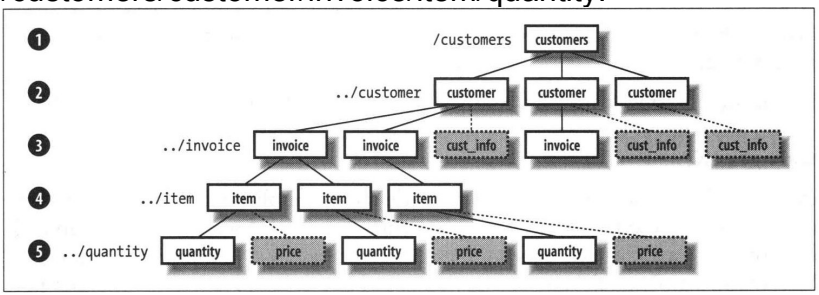

Figure 3: Selection of document nodes by a XPath expression [Sim02]

The fact that XPath is based on path expressions and the hierarchical structure of XML should not be taken to mean that XPath is a hierarchical query language. With predicates and support for wildcard expressions XPath is fully capable of declarative queries.

The XQuery Language

XQuery [RCF+06b] is a superset of XPath 2.0. In addition to the features of XPath, it also provides extensive support for document transformations and for the iterations over result sets. A central aspect of XQuery is the FLOWR acronym.

For The For clause is used for the iteration over the nodes of a resultset.

Let The Let clause assigns a set of node to a variable

Order , The Order clause allows to specify a node by which values the results are sorted.

Where The Where clause is used to restrict the results of a query expression with a condition
Return . The Return clause returns the current results.

The XQuery specification supports static typing, including the ability to check queries against the XML schema of the data queried. This ability can be used to detect errors in queries.

XML Updates

Neither XPath alone nor XQuery include support for document updates. For most XML databases, the creation or change of documents is either done with a proprietary language for one particular database or by using an API. An XQuery Update Extension [RCF06a] is currently being defined.

XML Document Collections

Besides the XML data model most XML databases support document collections. Each collection contains a number of documents where each document is a single tree of XML Elements. If a XML Database is compared to a file system, the documents would correspond to files and collections to directories. Databases not supporting collection view all data as a single XML tree, documents and collections are emulated by inserting nodes into this database tree.

Normalized Data an XML

The practice of data normalization is well established for relational databases, it is widely

researched and well supported, for example by the declaration of referential constraints. Data normalisation for XMLdatabases has to be handled differently since the basis XML data model is based on a tree, not a graph. The XML standard relevant in order to define references between different subtrees in XML Elements is XLink [DeR99]. XLink makes it possible to define manytomany links between document nodes without modifying the nodes themselves it it not necessary to include links in documents as it is the case with HTML. XLink is supported by some XML databases, especially by those with a background in the management of complex electronic publications. Figure 4 shows how XLink can be used to provide links to the editions of a book.

```
<novel xlink:type = "extended">
  <title>The Wonderful Wizard of Oz</title>
  <author>L. Frank Baum</author>
  <year>1900</year>
  <edition xlink:type="locator" xlink:href="urn:isbn:0688069444"
           xlink:title="William Morrow"
           xlink:role="http://purl.org/dc/elements/1.1/publisher"
           xlink:label="ISBN0688069444"/>
  <edition xlink:type="locator" xlink:href="urn:isbn:0192839306"
           xlink:title="Oxford University Press"
           xlink:role="http://purl.org/dc/elements/1.1/publisher"
           xlink:label="ISBN0192839306"/>
  <edition xlink:type="locator" xlink:href="urn:isbn:0700609857"
           xlink:title="University Press of Kansas"
           xlink:role="http://purl.org/dc/elements/1.1/publisher"
           xlink:label="ISBN0700609857"/>
</novel>
```

Figure 4: Use of XLink to provide ISBN references for the editions of a book [HM04]

XML Databases and Information Retrieval

The XQuery definition does not include full text retrieval nor language features which would allow to return documents ordered by their relevance to a given

query. The XQuery and XPath FullText [SBB+06] extension includes such features, but to the knowledge of the author no implementation exists yet. When implemented, the extension will provide many features useful for information retrieval on semi structured data, such as support for different ranking algorithms and mappings between thesauri.

Full Text Retrieval Architectures

Current fulltext retrieval systems like FAST Data Search or the Open Source library Lucene support retrieval over large datasets as well as fielded search. They therefore allow retrieval of semi structured data as well as of unstructured texts. Full text retrieval architectures have strong roots in information retrieval, they currently offer the best support for relevance measures. The following sections will explain aspects of the support that full text retrieval systems offer for semi structured and to a lesser degree structured data. The various aspects of documents used to calculate a relevance measure for a given user information need are introduced.

Support for semi structured Data

The most basic support for querying semi structured data is the ability of a retrieval system to treat a document as a collection of fields. This subdivision allows queries that not only take into account if a query value can be found in a document, but also where it can be found. This is useful for a number of reasons: It allows to limit queries on subsets of a document, for instance the title or the author, and also makes it possible to store metadata with documents. In addition documents divided into different fields can easily be mapped to rows in a

table in a relational database, thus allowing a certain degree of reuse for existing data schemas.

Besides the ability to subdivide documents into fields, support for different data types in addition to text is a significant feature of modern text retrieval systems. The data types most often supported are integers and dates. While both these data types can be expressed as text, storing them in their native format has a number of advantages.

. • Sorting on the data types becomes possible. When numbers and dates are treated as text, special formating rules must be followed in order to ensure the correct sort order, numbers for example must be prefixed with zeros else the number 88 will be sorted after the number 100.

. • Range queries can be supported. A Query for all documents published between 1969 and 1984 can only be correctly interpreted when the retrieval system is aware that fields are dates or numbers.

Some full text retrieval systems support structured documents in addition to fields, making it possible to search only for terms appearing in certain sections of a document. Figure 5 shows an document structured into chapters and sections, with the sections including a title. Support for such structured documents makes it possible to query XML documents without loss of structured information.

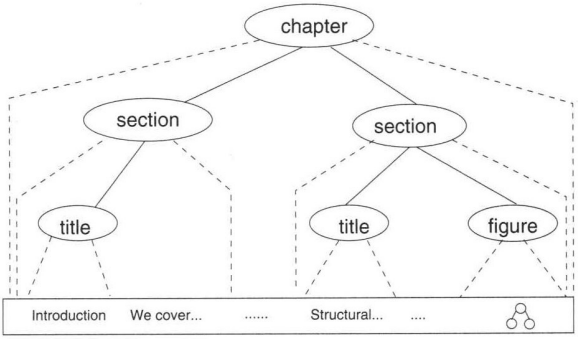

Figure 5: Logical representation of a structured document [BYRN99]

Measures of Relevance

In contrast to the boolean retrieval model where documents are either relevant or not relevant to a user's information need, modern information systems are based on statistical retrieval models where the degree of relevance of a document to a query can be expressed on a numerical scale. While the algorithms employed by today's commercial retrieval systems are in most cases a closely guarded secret of their vendor, the main characteristics of documents used for ranking are known.

A distinction can be made between topological ranking approaches such as the PageRank algorithm [PBMW98] and approaches based on the content of documents. Topological algorithms, while certainly relevant to the ranking of document collections including rich link relationships, are less relevant for document collections containing information about scientific research projects, it will therefore not be further discussed in this paper. Approaches based on the content of documents differ with system implementations, the FAST retrieval system for example offers to following ranking criteria:

Context The Context measure compares the query terms with the terms found in the documents, it is based on the TFIDF model [BYRN99].

Proximity Proximity takes the positions of terms in a document into account, the closer to each other terms occur in the document, the higher is the document ranked.

In addition to the measure listed above FAST does also provide the following ranking criteria not related to the content of the document:

Freshness Newer documents only recently added to the collection receive a higher index shore than older documents.

Quality A handicap or boost can be applied to the ranking value of a document. This makes it possible to let the importance of documents influence the ranking.

Roles and Use cases of the different Systems

After the brief introduction of the different approaches and capabilities of relational databases, XML data management systems and full text retrieval systems in the preceding section, this section describes some common use cases for information systems and highlights their respective strengths and weaknesses.

Data Retrieval

A data retrieval task is given if a definitive criteria exists to decide whether a particular dataset should be included in a result set or not. An example would be a query for all documents published in 1999 or all researchers in a database whose first name starts with the letter A. For such simple queries all three types of systems are well suited. The situation is more complicated when joins between di□erent datasets are

introduced. Full text Retrieval Systems do not o□er support for joins, the data would either have to be denormalized before storing it or two separate requests would have to be made. Both SQL and XQuery in contrast support joins. A possible problem for joins on XML databases is the fact that research and implementation of index structures for such operations is still in an early state. Finding the optimal index definition and query to perform a request may thus take some experimentation. Support for joins on relational dataset in contrast is well researched and documented.

Information Retrieval

An information retrieval task is given when the inclusion of a document in a query result and its position depend on the interpretation of an individual information need of a human user. In principle all three types of system are capable of information retrieval but full text retrieval systems are primarily developed for this task. Neither SQL nor XQuery do provide support for the ranking of documents by their relevance by default, though various extensions exist.

Transaction Handling

Both XML data management systems and relation database provide support for transactions. Transaction support for XML data management systems is often limited by the fact that locks are held on the level of documents since XML Documents are trees, each change in a tree potentially effects the whole tree. Full text retrieval systems do not provide support for transactions, even with modern systems supporting incremental changes of the index, often using a secondary index as shown in figure 6.

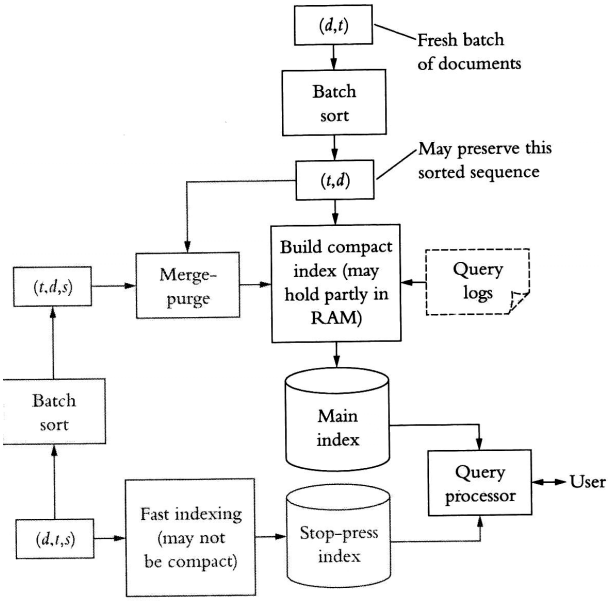

Figure 6: Use of two indices to for retrieval on dynamic data [Cha03]

Data Integration

The term data integration describes tasks where heterogeneous data schemas have to be integrated into a common schema in order to allow retrieval operations on the data. This data may come in the form of XML, text files or database schemas. The integration of different data schemas is a common problem in information processing, still, while considerable research was done, see [Ber03] for example, the application of the research results remains a challenge.

Relation databases are not well suited for the task of data integration. A relational database must have a schema at all times, it is therefore necessary to express all the different schemas to be integrated in the relational data model and then define one common schema which

is capable to express the concepts of all other schemas. In many cases such an approach is not practicable: Not only is the upfront design of a common schema a hard problem, but also do changes and requirements and the need to integrate further data collections at a later date necessitate changes to the relation schema.

Full text retrieval systems are also not suited for data integration task the focus of the systems on retrieval as opposed to write operations is problematic, as is the simple data model underlying most systems where documents are composed only of a set of flat fields without support for further structure or joins.

XML data management systems in comparison offer extremely rich support for data integration tasks. The first advantage XML data management systems have is the XML data model itself. Mapping data to XML from other data models is mostly trivial. This allows it to load data to be integrated early into a XML data management system the following steps of integration thus benefit from all advantages databases o□er. A second advantage of XML data management systems for data integration is the ability to begin integration without a schema and implement and refine a schema as the data integration task progresses. When the latter addition of new Elements is necessary, this element can be marked optional in the schema and only used for those documents that need them. Figure 7 shows the architecture of the Nimble [DHW01] System, using XML as a central data model to integrate data from various sources.

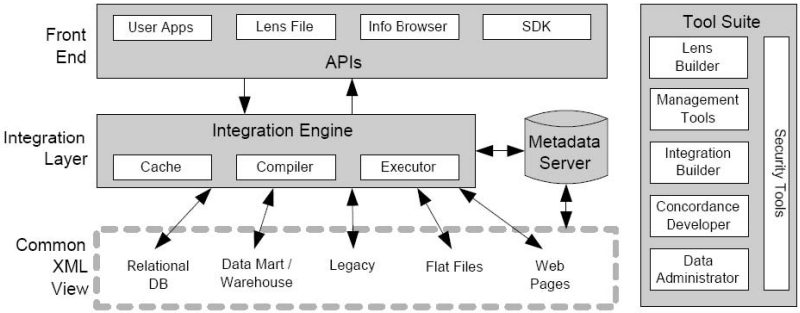

Figure 7: Architexture of the Nimble systems [DHW01]

Integration of the different types of Systems

As described in the preceding section, the different types of information systems have various strengths and weaknesses for the different use case. This leads to the question if the different systems can be integrated in order to supplement each other. Relational and XML databases provide many of the same feature: Both are databases with the relational databases better suited for highly structured data and the XML databases with better support for semi structured and hierarchical data, both commonly lack support for relevance ranking, thus the integration of a full text retrieval system and one of the database appears promising. The following sections will discuss how the data models and architectures of the different systems can be integrated.

Integration of relational databases and full text retrieval

The close resemblance between the fields of documents in full text retrieval systems and the values in the rows of relational databases ease the transfer from one system to the other. Since relational databases provide support for transaction and better means to

update existing data, we propose to use a full text retrieval system for information retrieval over the entries in the database. For most database schemas not all normalized relations will hold all information by themselves, therefore a view should be defined to include all values that should be searchable. Such a view would than be used as the data source for a full text information system. Depending on the amount of data in the database and the requirements for the freshness of the data, either incremental updates can be used or two full text retrieval systems are used in rotation, one providing search while the other is indexing the data.

Integration of XML data management systems and full text retrieval

Support for structured text retrieval is currently not a common feature of full text retrieval systems which makes the mapping from XML data management systems to full text retrieval systems without loss to structural information diffcult. One question that has to be answered when deciding on how to integrate the different data models is the degree of structural information relevant to the information need of a user, non relevant structural information can be discarded. When a XHTML document is made searchable for example, the H1 elements might be stored in a field mainheader, the elementsH2,H3,H4,... in the field coheader, the meta and title elements under the /head element might be stored under in the fields meta and title, respectively. All remaining content of the document might be stored under the field text.

When the above approach is not applicable because it would result in too many fields, it is possible to change the values in the fields by using a prefix. When the full text retrieval system only has a field person for

example, and information about the role of the person must be preserved it is possible to use a transformation as the following: <person role='author'> <firstname>Hans</firstname> <surname> Meier</surname></person> would became au:fn:Hans au:sn:Meier in the field person, where au would be short for author, fn for first name and sn for surname. Such an encoding would allow to query only for those documents where a person with the surname Meier was the author of the document. A disadvantage of such a solution would be a loss in performance: Not only do documents with prefixed fields take up more storage in the index, in some cases it would also be necessary to add a name to the index twice, once with and once without the index, with possible negative influences on ranking.

Since XML databases do not offer a view mechanism yet, regular queries would be used to obtain the data to index, XQuery would be well suited for this task since it includes the ability to transform results into an format suitable for indexing.

Conclusion

To repeat the question from the title of this text, where are future information architectures heading? Currently, information system technology is progressing fast, with each type of system gaining features that before were not available. On the level of the individual organisation, the choice for an information architecture is not only dependant on the system capabilities, but also on the already existing infrastructure. However, as the preceding section explained, combinations between the different types of systems are not only possible but for many reason beneficial. Where relational databases for instance provide solid support for transactions and data

retrieval, a full text retrieval system can act as a view on the data in the database, providing relevance ranking.

The social science information centre decided to use the XML data management system XHive and the full text retrieval system FAST ESP as the information architecture for the SOWIPORT portal. The reasons for this decision were the need to integrate collections of heterogeneous data provided by partners for inclusion in SOWIPORT. The integration of this data was done by incrementally developing a schema optimized for the retrieval of scientific information. While the focus of the schema was on bibliographic datasets and research projects, XML allowed enough flexibility to prepare the schema for further extensions. Beside its role in data integration, the XHive databases also allows for the formulation of complex request. Since these requests are written in XQuery, transformation of the data for further publication into multiple export formats is well supported.

While the XHive databases allow it to retain the same structured query capabilities currently implemented with a relational database, the FAST full text retrieval system is used to provide relevance ranked results for user information needs.

It is our belief that the combination of the relational CERIF schema with full text retrieval systems can provide a strong basis both for data processing and the fulfillment of user information needs.

References

[ABS00] Serge Abiteboul, Peter Bunemann, and Dan Suciu. Data on the Web: From Relations to Semistructured Data and XML. Morgan Kaufman, 2000.

[Ben03] Shawn E. Benham. XML Data Management, chapter IBM XMLEnabled Data Management Product Architecture and Technology, pages 91–129. Addison Wesley, 2003.

[Ber03] P. Bernstein. Applying model management to classical meta data problems, 2003.

[BYRN99] Ricardo A. BaezaYates and Berthier RibeiroNeto. Modern Information Retrieval. AddisonWesley Longman Publishing Co., Inc., Boston, MA, USA, 1999.

[CD99] James Clark and Steven DeRose. XML path language (XPath) version 1.0. W3C recommendation, W3C, November 1999. http://www.w3.org/TR/1999/RECxpath19991116.

[Cha03] Soumen Chakrabarti. mining the web: Discovering Knowledge from Hypertext Data. Morgan Kaufman, 2003.

[Cod70] E. F. Codd. A relational model of data for large shared data banks. Commun. ACM, 13(6):377–387, 1970.

[DeR99] Steven J. DeRose. XML XLink requirements version 1.0. W3C note, W3C, February 1999. http://www.w3.org/TR/1999/NOTExlinkreq19990224.

[DHW01] Denise Draper, Alon Y. Halevy, and Daniel S. Weld. The nimble XML data integration system. In ICDE, pages 155–160, 2001.

[FSC+06] Mary F. Fernandez, J´ome Sim´eon, Don Chamberlin, Michael Kay, Jonathan Robie, Scott Boag, and Anders Berglund. XML path language (XPath) 2.0. Candidate recommendation, W3C, June 2006. http://www.w3.org/TR/2006/CRxpath2020060608/.

[HM04] Elliotte Rusty Harold and W. Scott Means. XML in a Nutshell: A Desktop Quick Reference. O'Reilly, 2004.

[IDB+06] Masayasu Ishikawa, Micah Dubinko, Mark Birbeck, Beth Epperson, Shane McCarron, Ann Navarro, Jonny Axelsson, and Steven Pemberton. XHTMLTM 2.0. W3C working draft, W3C, July 2006. http://www.w3.org/TR/2006/WDxhtml220060726.

[KKN03] Rajasekar Krishnamurthy, Raghav Kaushik, and Je□rey F. Naughton. Xmltosql query translation literature: The state of the art and open problems. In Zohra Bellahsene, Akmal B. Chaudhri, Erhard Rahm, Michael Rys, and Rainer Unland, editors, Database and XML Technologies, Lecture Notes in Computer Science. Springer, 2003.

[LVLG03] Chengfei Liu, Millist W. Vincent, Jixue Liu, and Minyi Guo. A virtual xml database engine for relational databases. In Zohra Bellahsene, Akmal B. Chaudhri, Erhard Rahm, Michael Rys, and Rainer Unland, editors, Database and XML Technologies, Lecture Notes in Computer Science. Springer, 2003.

[Mei03] Wolfgang Meier. XML Data Management, chapter eXist Native XML Database, pages 43–67. Addison Wesley, 2003.

[PBMW98] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

[RCF06a] Jonathan Robie, Don Chamberlin, and Daniela Florescu. XQuery update facility. W3C working draft, W3C, July 2006. http://www.w3.org/TR/2006/WDxqupdate20060711/.

[RCF+06b] Jonathan Robie, Don Chamberlin, Daniela Florescu, Scott Boag, Mary F. Fernandez, and J´ome Sim´eon. XQuery 1.0: An XML query language. Candidate recommendation, W3C, June 2006. http://www.w3.org/TR/2006/CRxquery20060608/.

[RGM01] Sriram Raghavan and Hector GarciaMolina. Crawling the hidden web. In VLDB ’01: Proceedings of the 27th International Conference on Very Large Data Bases, pages 129–138, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[SBB+06] Jayavel Shanmugasundaram, Chavdar Botev, Stephen Buxton, Darin McBeath, Pat Case, Mary Holstege, Sihem AmerYahia, Jochen Doerre, and Michael Rys. XQuery 1.0 and XPath

2.0 fulltext. W3C working draft, W3C, May 2006. http://www.w3.org/TR/2006/WDxqueryfulltext20060501/.

[Sim02] John Simpson. XPath and XPointer. O’Reilly, 2002.

[Sne05] John Snelson. All xml databases are equal. In XTech 2005: XML, the Web and beyond, 2005.

[ST01] Airi Salminen and Frank Wm. Tompa. Requirements for xml document database systems. In DocEng ’01: Proceedings of the 2001 ACM Symposium on Document engineering, pages 85–94, New York, NY, USA, 2001. ACM Press.

[vdV02] Eric van der VList. XML Schema. O’Reilly, 2002.

[vdV03] Eric von der Vlist. Relax NG. O’Reilly, 2003.

Computational Data Mining for Automated Information Extraction from Biomedical Data Repositories

Kenneth Revett
Harrow School of Computer Science, University of Westminster, London, UK
revettk@westminster.ac.uk

Abstract

In this paper we describe computational approaches to extraction of information from biomedical data repositories, included in biomedical Current Research Information Systems. These data repositories vary in size (number of objects) and the amount of information (number of attributes) that they contain. They generally range from very small datasets with 100 objects and a few attributes to large DNA microarray datasets that contain thousands of attributes and only a few dozen examples for each of the attributes. This disparity in the size of biomedical data repositories necessitates a heterogeneous approach to automated information extraction. In this paper, we describe a rough sets approach to the automated extraction of information from such data repositories. Compared with other techniques – rough sets is robust enough to handle a variety of disparate information systems without the need for an extensive amount of data pre-processing.

Introduction

This paper present the results of a series of experiments where we have employed rough sets to classify small biomedical datasets containing information on several major diseases (diabetes, heart disease and cancer). Medical datasets tend to be difficult to obtain, as highly skilled professional effort is required to produce

the information. These datasets can be small or large, contain missing values, with mixed value types for the attributes (i.e. nominal versus continuous) and hence may require special treatment if they are to be mined for their information content. Small biomedical datasets are somewhat difficult to extract information from because they may not contain sufficient information to represent the domain sufficiently. Large datasets tend to be overly sparse, such as those found in most typical microarray datatsets – which may contain 10,000s of attributes (genes) and only a small number (typically 20-40) examples. Each scenario presents difficulties when attempting to extract useful information.

Generally one mines biomedical datasets in order to extract some causal relationship between the attributes and the decision class (which will usually be the presence/absence of a given medical condition). If the results are to be useful to the clinician, the relationship should be in the form of a readily interpretable rule. These are several such rules types: characteristic, association, and decision rules. These approaches can be implemented on a data cube method or an attribute-orientated induction method. Characteristic rules are a conjunction of properties shared by all entities within a class. In contrast, discriminant/decision rules are conjunctions of properties shared by all entities in the target class and distinguishes its entities from contrasting classes. These approaches, which are essentially concept-driven differ in significant ways from the on-line analytical processing (OLAP) approach. First, the OLAP approach is user-centric – in that it requires users to initiate the information extraction process. Processes such as roll-up, dicing, and slicing must be initiated by the user. In addition, OLAP systems

assume dimensions to be categorical and operate on numerical data only. These restrictions do not apply to characteristic rules – nor do they apply to decision rules such as those that are generated from a rough sets approach.

In the next section, we briefly introduce the concepts of rough sets, followed by the results of a series of examples where we have applied rough sets. The examples presented in this paper concern small biomedical datasets. A brief discussion of the application to large datasets (such as microarrays) will be left to the discussion section. Lastly we summarise the results of this study and make suggestions regarding the how medical professionals can increase the amount of potential information content of their datasets.

Rough Set Theory

Rough sets were first reported in the literature by prof Pawlak in the early 1980s [11]. The first step in the process of mining any dataset using rough sets is to transform the data into a decision table. In a decision table (DT), each row consists of an observation (also called an object) and each column is an attribute, and a decision attribute for the observation {d}. Formally, a DT is a pair A = (*U,* A∪{d}) where d ϖ A is the *decision attribute, U* is a finite non-empty set of objects called the *universe* and A is a finite non-empty set of attributes such that $a:U->V_a$ is called the value set of a. Once the DT has been produced, the next stage entails cleansing the data.

There are several issues involved in small datasets – such as missing values, various types of data (categorical, nominal and interval) and multiple decision classes. Each of these potential problems must be

addressed in order to maximise the information gain from a DT. Missing values is very often a problem in biomedical datasets and can arise in 2 different ways. It may be that an omission of a value for 1 or more subject was intentionally – there was no reason to collect that measurement for this particular subject (i.e. 'not applicable' as opposed to 'not recorded'). In the second case, data was not available for a particular subject and therefore was omitted from the table. We have 2 options available to us: remove the incomplete records from the DT or try to estimate what the missing value(s) should be. The first method is obviously the simplest, but we may not be able to afford removing records if the DT is small to begin with – the focus of this paper. So we must derive some method for filling in missing data without biasing the DT. In many cases, an expert with the appropriate domain knowledge may provide assistance in determining what the missing value should be – or else is able to provide feedback on the estimation generated by the data collector. In our studies, we employ 1 of several options to fill in missing data items: i) mean/mode fill or ii) conditioned mean/mode fill. In each case, the mean or mode is used (in the event of a tie in the mode version, a random selection is used) to fill in the missing values, based on the particular attribute in question. The difference between the two methods is whether the decision attribute partitions the attribute values for which the mean is taken. Any other method that is suitable for the data at hand will suffice (see [7,12] for discussions on this issues). Once missing values are handled, the next step is to discretise the dataset.

Rarely is the data contained within a DT all of ordinal type – they generally are composed of a mixture of ordinal and interval data. Discretisation refers to

partitioning attributes into intervals – tantamount to searching for “cuts” that determine the intervals. All values that lie within a given range are mapped onto the same value, transforming interval into categorical data. In general, there are three basic ways of performing discretisation:

Each attribute is considered in isolation
Each attribute is selected, conditioned on the decision value
All attributes are considered simultaneously, conditioned on the decision value(s)

As an example of a discretisation technique, one can apply equal frequency binning, where a number of bins *n* is selected and after examining the histogram of each attribute, n-1 cuts are generated so that there is approximately the same number of items in each bin. See the discussion in [12] for details on this and other methods of discretisation that have been successfully applied in rough sets. Now that the DT has been pre-processed, the rough sets algorithm can be applied to the DT for the purposes of supervised classification.

The basic philosophy of rough sets is to reduce the elements (attributes) in a DT based on the information content of each attribute or collection of attributes (objects) such that the there is a mapping between similar objects and a corresponding decision class. In general, not all of the information contained in a DT is required: many of the attributes may be redundant in the sense that they do not directly influence which decision class a particular object belongs to. One of the primary goals of rough sets is to eliminate attributes that are redundant – a reduction task. Rough sets use the

notion of the lower and upper approximation of sets in order to generate decision boundaries that are employed to classify objects. Consider a decision table A = (U, A∪{d}) and let $B \subseteq A$ and X ⊆ U. What we wish to do is to approximate X by the information contained in B by constructing the B-lower (B_L) and B-upper (B^U) approximation of X. The objects in B_L ($B_L X$) can be classified with certainty as members of X, while objects in B^U are not guaranteed to be members of X. The difference between the 2 approximations - B^U - B_L, determines whether the set is rough or not: if it is empty, the set is crisp otherwise it is a *rough set*. What we wish to do then is to partition the objects in the DT such that objects that are similar to one another (via some similarity metric) are treated as a single entity. This collection of similar entities is considered consistent *iff* they all map to the same decision value. Unfortunately in many Biomedical datasets, many objects are inconsistent, which tends to reduce the classification accuracy of the rough sets algorithm unless this situation is handled appropriately (see [6,13,15,16] for a discussion of this important topic). The next step is to reduce the DT to a collection of attributes/values that maximises the information content of the decision table. This step is accomplished through the use of the indiscernibility relation *IND(B)* and is defined for any subset $B \subseteq A$ ($B \subseteq A \cup \{d\}$) as follows:

$$IND(B) = \left\{ (x, y) \in U \times U : \text{for every } a \in B \; a(x) = a(y) \right\} \quad (1)$$

The elements of *IND(B)* correspond to the notion of an equivalence class. The advantage of this process is that any member of the equivalence class can be used to represent the entire class – thereby reducing the

dimensionality of the objects in the DT. This leads directly into the concept of a *reduct*, which is the minimal set of attributes from a DT that preserves the equivalence relation between conditioned attributes and decision values. It is the minimal amount of information required to distinguish objects with in U. The collection of all reducts that together provide classification of all objects in the DT is called the *CORE*(A). The CORE specifies the minimal set of elements/values in the DT which are required to correctly classify objects in the DT. Removing any element from this set reduces the classification accuracy. It should be noted that searching for minimal reducts is an NP-hard problem, but fortunately there are good heuristics that can compute a sufficient amount of reducts in reasonable time to be usable. In the software system that we employ an order based genetic algorithm (o-GA) which is used to search through the decision table for approximate reducts [2,14, 16]. The reducts are approximate because we do not perform an exhaustive search via the o-GA which may miss one or more attributes that should be included as a reduct. Once we have our set of reducts, we are ready to produce a set of rules that will form the basis for object classification.

Rough sets generates a collection of 'if..then..' decision rules that are used to classify the objects in the DT. These rules are generated from the application of reducts to the decision table, looking for instances where the conditionals match those contained in the set of reducts and reading off the values from the DT. If the data is consistent, then all objects with the same conditional values as those found in a particular reduct will always map to the same decision value. In many cases though, the DT is not consistent, and instead we

must contend with some amount of indeterminism. In this case, a decision has to be made regarding which decision class should be used when there are more than 1 matching conditioned attribute values. Simple voting may work in many cases, where votes are cast in proportion to the support of the particular class of objects. In addition to inconsistencies within the data, the primary challenge in inducing rules from decision tables is in the determination of which attributes should be included in the conditional part of the rule [4,9,10]. If the rules are too detailed (i.e. they incorporate a maximal number of attributes), they will tend to overfit the training set and classify weakly on test cases. What are generally sought in this regard are rules that possess low cardinality, as this makes the rules more generally applicable. This idea is analogous to the building block hypothesis used in genetics algorithms, where we wish to select for highly accurate and low defining length gene segments [2]. There are many variations on rule generation, which are implemented through the formation of alternative reducts such as *dynamic* and *approximate* reducts. Discussion of these ideas is beyond the scope of this paper and the interested reader is directed towards [1,8] for a detailed discussion of these alternatives.

The rules that are generated are in the traditional conjunctive normal form and are easily applied to the objects in the DT. What we are interested in is the accuracy of the classification process – how well has the training rule set classified new objects? In addition, what sort of confidence do we have in the resulting classification of a particular validation training set? These are standard issues that hold true for any machine learning application. In addition questions arise regarding methods for handling biomedical datasets that contain an

unequal distribution of decision class objects. Traditionally in rough sets, validation is accomplished through N-fold validation, where the N is dependent upon the particular dataset at hand – but generally a 70/30 training/validation scheme is used, with replication with replacement on the order of 10% of the sample size.

In the next section, we present some results from several studies of small biomedical datasets where we have applied rough sets. We only briefly describe the results, focusing on the factors that influence the classification accuracy, dimensionality reduction and rule generation facility that is inherent to the rough set paradigm.

Results

In this section, we report results obtained from applying rough sets as a classification tool on a collection of relatively small biomedical datasets. We report findings related to pre-processing strategies, dimensionality reduction, classification accuracy and rule generation. The aim is to demonstrate the robust applicability of rough sets both in regards to the types of datasets one can examine.

Pre-processing techniques:

Most datasets require some form of pre-processing. The amount of pre-processing required is dependent on the type of data and the care taken to generate the data. In the context of small biomedical datasets, the clinician(s) involved in generating the data may or may not be conversant with standard data mining techniques. Obviously, the more informed the clinician(s), the more likely the data will be in the appropriate form for datamining. In Table 1 below, we present a section of a

decision table that required pre-processing in the form of filling in missing values and discretisation. The dataset contains instances of prostate cancer with 502 objects, 18 attributes and 10 decision classes.

String	Integer	Integer	Integer	String	Integer	Integer	Integer	String
0.2 mg estrogen	72	75	76	normal activity	0	15	9	heart strain
0.2 mg estrogen	1	?	116	normal activity	0	13	7	heart block or conduction def
5.0 mg estrogen	40	69	102	normal activity	1	14	8	heart strain
0.2 mg estrogen	20	75	94	in bed < 50% daytime	?	14	7	benign
placebo	65	67	99	normal activity	0	17	10	normal
0.2 mg estrogen	24	71	98	normal activity	0	19	10	normal

Table 1: A section containing 6 objects (only 9 attributes are displayed to conserve space) from a prostate cancer dataset, available from the internet at: http://biostat.mc.vanderbilt.edu/twiki/pub/Main/Datasets.Note that the '?' in the column labelled 'age' and 'hx' indicate missing values.

This dataset is of moderate size with 502 objects, but contains a mixture of attribute types (integer and String) and different levels of measurement (interval and categorical). There are few techniques that allow direct classification of a dataset with this variety of measurement types and values. With rough sets, datasets of this type can be directly processed – albeit with reduced classification accuracy (we obtained 60% classification accuracy without any pre-processing). We subsequently pre-processed the dataset according to the following scheme:

filled in missing values using mean/mode fill algorithm conditioned on the decision

discretised all interval attributes using equal frequency binning (using 3 ranges corresponding to low-middle-high)

combining the 'dead' decision classes (9) into a single category.

The last pre-processing step was necessary because many of the decision classes were under-represented and therefore did not provide sufficient training and testing examples. With these pre-processing steps, we were able to achieve a classification accuracy using 300 cases (60%) for training and 202 for testing of 92% (maximal value over 20 randomised classifications – range 86%-92%).

Rule generation

The production of classification rules is probably the most striking feature of rough sets. Rules are in the form of if (attribute A = X AND attribute B = Y) then (decision => Z). These rules are simple to understand and interpret – something that is difficult for most techniques such as neural networks, genetics algorithms etc to perform. Decision tree algorithms have this facility, but lack some of the robustness of the rough set approach even with considerable pre-processing.

The important considerations with respect to rule generation is the ability to interpret the rules and the number of rules generated. The number of rules is obviously dependent upon the number of attributes, but is also critically dependent upon the level of discretisation. Please see [3] for a detailed discussion of these issues in a study of the Pima Indian Diabetes dataset.

Conclusion

The purpose of this paper was to introduce to the actors of biomedical CRIS, both medical practitioner/researcher and data-miner alike how one can

begin to apply rough sets to a typical biomedical dataset. As with most machine learning applications, pre-processing of the dataset is mandatory. It is hoped that through an understanding of the process of mining these types of datasets, medical practitioners and members of the allied health field in general will become more aware of the requirements for processing biomedical datasets. The primary considerations discussed in this paper were:

minimal number of missing values
consistency between attribute measurements (ordinal versus categorical)
having a balanced and sufficient number of objects for each decision class

When it comes to large datasets – such as those typically generated by microarray analyses, additional caveats become manifest. For one, most microarray datasets contain many more attributes – typically tens-of-thousands than they do class objects. This typically requires the use of significant data reduction techniques before rule extraction. Generally, the principal pre-processing step entails a thresholding process that culls out all non-call spots. This typically reduces the number of candidate attributes (genes) to tens-to-hundreds – a much more manageable set of attributes. In addition, many authors employ a clustering approach to remove non-significant genes. Even with these processing steps – the number of attributes generally outweighs the number of observations. Further analysis with rough sets has yielded very promising results [5,6,7,17]. None-the-less – this is still a very active area of research.

With these caveats in mind, KDD tools such as rough sets can be successfully applied in many cases.

Rough sets provide the ability to compensate for these difficulties that often occur in small/sparse biomedical datasets. In addition, they are able to generate highly accurate classifiers via a set of easy to understand rules. The combination of all of these features makes rough sets a very promising technique for the analysis and knowledge discovery process in small as well as large biomedical datasets.

Appendix A: A partial listing of software resources for implementations of rough sets:

Rosetta:
http://www.idt.unit.no/~aleks/rosetta/rosetta.html
Datalogic/R:
http://ourworld.compuserve.com/homepages/reduct
Reduct & Lobbe Technologies:
http://www.reduct.com/

References

Bazan, J.G., Skowron, A. & Synak, Piotr. Dynamic reducts as a Tool for Extracting Laws from Decision tables, Proceeding of the Third International Workshop on Rough Sets and Soft Computing, San Jose, California, pp 526-533, 1994.

Goldberg, D.E. *GA in Search, Optimisation, and Machine Learning*. Addison-Wesley, 1989.

Khan, A & Revett, K. Data mining the PIMA Indian diabetes database using Rough Set theory with a special emphasis on rule reduction, INMIC2004, Lahore Pakistan, pp. 334-339, December, 2004.

Klosgen W. & Zytkow J.M. (Eds) Handbook of data Mining and Knowledge Discovery, Oxford University Press, 2002.

Komorowski, J., Pawlak, Z., Polkowski, L & Skowron A. (eds): Rough Fuzzy Hybridization a New Trend in Decision Making. Springer Verlag pp. 3–98, 1999.

Komorowski, J., Hvidsten, T.R., Jenssen, T-K., Tieldvoll, D., Hovig, E., Sandvik, A.K., & Laegreid, A., Towards Knowledge Discovery from cDNA Microarray Gene Expression Data. PKKD 2000, pp. 470-475

Midelfart, H., Komorowski, J., Nørsett , K., Yadetie, F., Sandvik, A.K., & Lægreid, A. Learning rough set classifiers from gene expressions and clinical data Source Fundamenta Informaticae,Vol 53(2), pp 155 – 183, 2002

Nguyen H.S. & Skowron, A. Quantization of real-valued attributes, proc Second International Conference on Information Science, pp 34-37, 1995.

Nguyen, S.H., Polkowski, L., Skowron, A., Synak, P. & Wróblewski, J., 1996. *Searching of Approximate Description of Decision Classes*. Proc. of The Fourth International Workshop on Rough Sets, Fuzzy Sets and Machine Discovery, RSFD'96, Tokyo, November 6-8, pp 153-161, 1996.

Øhrn, A. "Discernibility and Rough Sets in Medicine" Tools and Applications. Department of Computer and Information Science. Trondheim, Norway, Norwegian University of Science and Technology: 239, 1999.

Pawlak, Z. Rough Sets, International Journal of Computer and Information Sciences, 11, pp. 341-356, 1982.

Pawlak, Z.: Rough sets – Theoretical aspects of reasoning about data. Kluwer Publishers, 1991.

Revett, K. & Khan, A. Rough Sets Based Cancer Classification System, IADIS 2005, pp 194 – 199.

Slezak, D.: Approximate Entropy Reducts. Fundamenta Informaticae, 2002.

Slezak, D., Wroblewski, J.: Order-based genetic algorithms for the search of approximate entropy reducts. In: Proc. of RSFDGrC'2003. Chongqing, China, 2003.

Wroblewski, J.: Theoretical Foundations of Order-Based Genetic Algorithms. Fundamenta Informaticae 28(3-4) pp. 423–430, 1996.

Ziarko, W. The Discovery, Analysis and Representation of Data Dependencies in Databases. In Piatesky-Shapiro, G. and Frawley, W.J. (eds.) Knowledge Discovery in Databases, AAAI Press/MIT Press, 1991, pp. 177-195.

Data retrieval in the PURE CRIS project at 9 universities – A practical approach[1]

Bo Alrø
Atira A/S
Denmark
ba@atira.dk

Introduction

Intro

This paper will describe the full set of data gathering and other retrieval methods for the PURE CRIS system. This set comprises methods for:

- Validated manual data entry
- Dynamic integration to local back-end systems
- Structured aggregation, enrichment and import of data from known sources
- Initial experiments with automated imports from known sources

Further, equal focus was placed on openly exposing data in corresponding manners:

- Web services (RPC/encoded + doc./literal) with rich methods
- Generic XML I/O
- OAI-PMH based data-providing

1 Based on an original paper by Aalborg University Library, Thøger Kristensen, Head of Acquisitions tk@aub.aau.dk and Atira A/S, Bo Alrø, Product Manager, ba@atira.dk.

- Z39.50 search support for library systems (SRW/SRU[1])
- Reporting engine (.doc, .pdf, .xml, .rtf, .csv)
- Generic XML I/O
- PUREportal framework

Data exposure will also be discussed later.

Finally, in some user cases, much of the data gathering and other retrieval effort aims at long term archiving. For that reason integration to DSpace and FEDORA will be briefly mentioned, too.

Retrieval

In this section, the full set of data gathering and other retrieval methods for the PURE CRIS system is described.

Validated manual data gathering

The basics of data gathering at universities using PURE is manual data entry. For this process, PURE supports de-centralized data submission from the entire academic body.

The idea is to ease the workload on the central library professionals and in the same time have the researchers and the research staffs at each institutes take an active role and interest in archiving and exhibiting the institution's academic production.

Distributed user rights, roles, and workflow

This de-centralized data submission is based on individual roles per user. This means, that the individual user operates the system and carries out tasks within a

[1] Search and Retrieve Web Service and Search and Retrieve URL Service are to be implemented later

limited role. The role defines the user's rights. A user profile can well consist of several roles at the same time.

Users access the user-interface of PURE from within a regular internet browser. This combined with the concept of roles allows for the browser-interface to be adapted precisely to the role: No user is exposed to more features than he or she actually needs. Also, the user-interface has built-in help and local vocabulary to further support the user.

Finally when looking at data gathering, manual data submission is tied into a workflow, which enables peer-validation of submitted data to ensure both completion and quality before data finally are committed to the repository.

The point here is not the technical specifications but the fact that the entire task of ongoing manual data entry throughout the academic year is done de-centrally at the individual faculties, institutes and departments. This rallies for a more effective and dynamic organisational processes, compared to central gathering, and it creates value in other ways, too:

1. Researchers get involved: A) instant searchability internally, only ca. 1-2 weeks on Google. B) Dynamic benchmarking throughout the academic year. C) Increased citations[1].
2. Library staff is relieved and consequently engaged in more value creating activities than central data entry.

[1] Indicated by these studies: "The Citation Impact of Digital Preprint Archives for Solar Physics Papers", Travis S. Metcalfe, 2006 (http://www.arxiv.org/abs/astro-ph/0607079) and "The Effect of Use and Access on Citations", Michael J. Kurtz et al, 2005 (http://www.arxiv.org/abs/cs/0503029). The effect has been questioned in smaller scientific areas, however.

3. University management at all levels (University, Faculty, Institute) gets concurrent reporting throughout the academic year.
4. Repository accumulates (richer) content concurrently rather than annually.

Dynamic integration to local back-end systems

A large portion of data in a fully ingested PURE repository derives from a number of local back-end databases at the research institution itself. That is systems such as personnel systems, financial packages and other administrative and technical systems like LDAP, Active Directories or SSO- and access security systems.

The classical data objects in the repository, which derives in part or in full from such systems, are:

1. Persons (and their roles)
2. Projects (research projects, registration and administrative data)
3. Organisations (faculties, institutes, departments)

Also secondary content types such as news clippings or bibliometrical data (usually citations and impact factors) can derive from such systems.

Most of the PURE-using research institution demand the ability to dynamically integrate PURE to such systems. Therefore, PURE offers a technical framework with a plugin architecture for customizing integration to any such source.

Integration to this type of systems are usually set up as logically controlled synchronisations[1] based on batch runs. In that case, data is kept redundantly in the repository. Alternatively, integration can also be set up directly to the source system in real-time, depending on

[1] Data integrity and versions are controlled and verified per run

the local system in question. However, the demand for such real-time integration have been very limited.

Aggregation, enrichment and import of historic data

Definition of historic data

By historic data we mean publication data or other IR-relevant data which is no longer updated, but which is wanted in an institutional repository. An example could be publication registrations from previous years.

When research institution implements a new PURE-based institutional repository, aggregation, enrichment and import of such historic data is very often part of the project scope.

The PXA format

For the purpose of importing data into PURE, there is a defined data import format called PXA (PURE XML Archive format). It is a zip-based container format, which handles both meta-data, binary files, and relations between all object types.

A valid PXA-files must be created outside of PURE. If valid, it will import valid, interrelated objects directly into the repository. For creating valid PXA-files, tools and documentation exist. These are at present used exclusively by the application development team. A recent decision aims at making tools and documentation available to local developers and 3rd party consultancy- and service providers as demand arise.

Converting data from any given source to PXA can involve more or less manual work, depending on the source data. An good example is, that sufficient data about relations between objects - as defined by the ruling meta-data model in PURE - need to be present in the import data set in order to generate correct relations. Are

such information not available, a number of manual and semi-manual processes can be brought into play for enriching data accordingly.

Using the PXA format for importing historic data into PURE can be seen as a a set of processes, out of which many are manual.

Experiments with automated imports of historic data

Based on data format specifications from known sources[1], options for tool-based import into PURE is being investigated. A tool-per-source approach is followed.

Depending on data quality, such tool-based imports can be more or less automated. So far, experiments with import tools have been carried out for sources such as PubMed, ISI, Scopus, CT.gov, CSR.org, DDF (Danish), and a few local but large sources at leading Danish universities.

Exposure

All data from a PURE repository are automatically exposed in a number of interfaces in order to facilitate easy retrieval from the PURE repository.

Two web services

Two web services exits in PURE: RPC/encoded and Document/Literal, each with a rich library of methods.

Both these web services a standard XML over SOAP with WSDL-file documentation. What separates them from the usual services are the very rich libraries of methods that comes with each service. These methods

[1] In cases of lacking data format documentation, own analysis of data structures have been carried out

accurately reflects the meta-data model in use[1], thus offering a direct way of creating the most wanted listings and displays on any web and/or CMS platform in a matter of minutes.

Among other options, data can be exhibited directly from these web services in library formats such as APA, VANCOUVER, HARVARD and others. Usually, data is separated from representation, of course. But in the particular case of these library formats, offering this option saves man-hours for the web-side implementation team.

ASP code example

```
// Obtain a reference to the DLWS stub
     pure.PureWebService_session ws = (pure.PureWebService_session)Session["ws"];
     if (ws == null)
     {
         ws = new pure.PureWebService_session();
         ws.CookieContainer = new System.Net.CookieContainer();
         Session.Add("ws", ws);
     }

     // Get a maximum of 20 publications offset 10 positions from the beginning
     pure.GetPublicationsRequestType pubReq =
         new pure.GetPublicationsRequestType();
     pubReq.offset = "10";
     pubReq.size = "20";
     pubReq.light = true;
     pure.WsContentListResultType pubs = ws.getPublications(pubReq);
     Response.Write("Got " + pubs.result.Length +
         " (complete count: " + pubs.count + ")<br />\n");

     // Iterate publications and output their titles
     foreach (pure.PublicationType p in pubs.result)
     {
         Response.Write(p.title_pri + "<br />\n");
     }
```

OAI

OAI-PMH based data-providing for structured meta-data model harvesting is available.Currently, three formats are supported under the OAI-PMH protocol in

[1] Usually the PURE default meta-data model, but any meta-data model can be implemented

PURE: Dublin Core, the Danish national DDF-MXF format and the Swedish national SVEP format.[1]

A technical framework in PURE allows for easy implementation of further formats under OAI-PMH.

Z39.50

Z39.50 is currently supported, allowing e.g. the integration of the university repository in the Metalib platform of the university library. SRW/SRU will be supported at a later time[2].

Reports

A reporting engine with rich GUI contains a number of ready-made reports for listing and calculating data in PURE. Each report can be customized, and new custom reports can be defined, too. A similar set exits for bibliometrical reporting and statistics.

Reports can be rendered on-the-fly as .txt, .pdf, .xml, .rtf, .csv. and html.

Reports have recently become of particular importance to research institution managers at all levels. In Denmark, due to the fact that most universities use the PURE meta-data model, reporting and benchmarking across universities have become possible. Now, work is in progress to define a new national standard meta-data model for universities and the like, It will primarily be based on the PURE meta-data model.

Reference Manager

File export of specified data from PUER is possible directly in a native reference Manager format.

[1] Currently under implementation

[2] Search and Retrieve Web Service and Search and Retrieve URL Service are to be implemented later

The underlying proposition to researchers is, that if they register their data in PURE, they will not need to register twice in order to concurrently maintain their reference database.

Portal framework

PUREportal is a development framework used for rapid development and deployment of web portals, which automatically displays data from a PURE repository, thus reflection the meta-data model in use and its relations in full detail.

Such PUREportal-based portals are not to be taken for CMS systems. There are no article editors, and manually added articles are not an option. Rather, such portals should be seen as cost-effective alternatives. There is little need for requirement specifications and no need for data modelling in regard to web exhibition.

For large research institution, this type of portal offers a fast track to web publishing

and into the Google indexes.

Archiving

Finally, in some user cases, much of the data gathering and other retrieval effort aims at long term archiving. In matters of archiving, there is a particular focus on the full text file.

PURE stores meta-data in a customer-definable SQL environment through a Hibernate-facilitated O/R mapping layer.

Binary objects (e.g. full text publications) are stored in the file system of the servers running PURE. The idea was to keep it simple within the PURE scope and leave advanced full-text storage to systems already specialized in that discipline.

For that reason, ready-made connectors to DSpace and FEDORA has been developed, thus facilitating rich binary storage to LTP systems in a lean manner.

From the pure website at http://pure.atira.dk: "Both connectors are developed by Atira A/S. No license payment is charges for either connector. The Fedora-specific connector was developed in a DEFF-project headed by Mr. Mogens Sandfaer, DTV, and this connector is also Open Source. Development of both connectors was concluded by fall, 2005."

Next step

The near future regarding data retrieval will be dominated by initiatives like:

- More automated imports using increasingly advanced converters
- Automated data delivery (push and harvest) to industry specific search services (such ad PubMed, Nordicom), Documentary data collections (such as clinicaltrials.org), and national collections (such as DDF (DK), ForskDok (NO), etc.
- Temporary import objects

Temporary import objects

This object type is to be used when importing data from external sources, which is not sufficiently rich for creating a whole object (e.g. a "Person") or not in sufficiently quality to create a valid object.

Either of the to situations are often the case when importing from external sources. In many cases because the meta-data models used in PURE are very detailed and therefore very demanding. In other cases the

originating data source is in poor quality. An example would be PubMed, where many fields are "free text".

Another situation is when data cannot be properly related to other objects upon import.

In all these situations, where complete or valid objects cannot be created upon import of external data, temporary objects can be created instead. Next, additional functionality and related GUI will allow entitled users to correct or enrich any temporary object and then to save it as a complete or valid object.

Author Index

IWIRCRIS 2006 is a workshop that aims to address the main issues concerning information retrieval (IR) in the context of the Current Research Information Systems (CRIS). The existence of special needs resultant from the special characteristics of a CRIS was acknowledged in the 8th International Conference on Current Research Information Systems held in Bergen, Norway.

In CRIS-IR 2006, researchers from the IR field are invited to reflect upon the special issues raised by CRIS and to present their results in a workshop oriented for both technical and managing people.

ISBN: 978-972-98921-6-5