Assisting the Discovery and Reuse of Document-based Knowledge using Semantic Metadata

Hinnerk Brügmann

Department of Information Systems, University of Erlangen-Nuremberg
Lange Gasse 20, 90403 Nuremberg, Germany
hinnerk.bruegmann@wiso.uni-erlangen.de

1 Introduction

Due to the high importance computer-mediated communication and IT systems have in the execution of today's business processes, enterprises are producing a rapidly growing amount of information as by-product of their business activities (Zadok et al. 2004, p. 2f). Lyman and Varian (2003, p. 4) estimate the annual growth of newly created information to be 30%.

According to studies of Merrill Lynch and Ferris Research only 20% - 40% of this information is stored in a structured, semantically described form in electronic databases or structured business applications (e.g. SAP R/3). The other 60% - 80% are contained in the unstructured form of electronic documents[1] (e.g. Microsoft Office files, emails, images, multimedia files, or web based content). This is even more dramatic as the annual growth rate of newly created unstructured content is considerably higher than that of structured information. (Ferris Research 2008; Blumberg and Atre 2003)

The unstructured nature of the bulk of existing as well as newly generated information is cause for a whole host of inefficiencies and problems. In a single-user desktop environment unstructured documents are spread across local folder hierarchies and contained as attachments in email messages. This leads to potential redundancy and an increased required effort to locate document-based knowledge. Such problems multiply in a multi-user enterprise environment with numerous competing repositories, document management systems, shared network drives and local file systems for each distinctive user. On one hand a person's ability to overlook the available number of information sources diminishes as the amount of documents in numerous different repositories grows. On the other hand a multi-

[1] In this context the term document is understood as relating to electronic files containing unstructured information.

user environment with local file system repositories lacks the central coordinating instance of a single-user environment (where such would be the individual user himself).

As a result individual knowledge workers in such multi-user enterprise environments spend a substantial part of their working hours looking for the correct and most up-to-date information needed in workflow steps or tasks (Guther 2007, p. 25).

On an enterprise wide level there is one scenario we want to point out as an exemplary motivational use case which is usually referred to as E-Discovery in the context of legal litigation (Runyon 2007, p. 3f). Lawyers and internal staff of large institutional clients face the problem of how to efficiently conduct searches for relevant documents in large heterogeneous electronic data sets, for the purpose of responding to litigation demands. According to Baron and Thomson lawyers typically overestimate their true rate of recall, i.e. how well their searches for documents have uncovered all relevant evidence (Baron and Thompson 2007, p. 141). For this reason the analyst firm Forrester Research expects that companies will heavily invest in technology and software products to assist in the discovery process. But even though they estimate spendings to grow from $1.4 billion in 2006 to more than $4.8 billion in 2011 the methods employed by today's professional E-Discovery tools still require a high amount of additional reviewing to validate the results (Murphy 2006, p. 3).

While investments in advanced Enterprise Content Management (ECM) solutions are made these often fail to meet expectations to bring order into the document chaos as business units and individual users tend to circumvent seemingly cumbersome ECM restrictions on document handling (Gilbert et al. 2006, p. 3f). On top of that a working Information Lifecycle Management, which could at least partly alleviate the problem of information overload and redundancy by deleting outdated or redundant documents, is missing in many enterprises.

This paper illustrates an approach to assist in the enterprise wide lifecycle management and discovery of electronic documents by interrelating unstructured documents and correlating those documents to known business entities (employees, partners, projects, products or processes) in a querieable way.

2 Related Work

Previous research approaches to establish and make use of relationships among documents can be separated into two groups. The first group makes use of the content of documents to detect similarity or references in documents, while the second group of work relies on analyzing user activity involving documents usage to infer document interrelations.

2.1 Content-based Relation Building

Implicit structure and similarity in document content can be used to define and measure the relationship between two documents.

The project Stuff I've Seen (Dumais et al. 2003) relies on document metadata and in part document content to create a context-snapshot of the time when a document was accessed. A user might then at a later time query the Stuff I've Seen data-store with arbitrary keywords to be presented with a list of documents he accessed earlier. The project has a strong emphasis on the user-interface part to allow for convenient querying those relations during a users day-to-day work in information-heavy workflow tasks. This interaction allows the user to quickly refine their query based on whatever contextual knowledge he can remember. A second research project, Haystack (Adar et al. 1999) consolidates all accessed information in its own internal information-store and tries to recognize the occurrence of named entities (e.g. persons) in texts. In the Haystack data model, a typical application file is shredded into many individual information objects of various types that are connected through application-specific relationships. A main prerequisite for such an approach is the need to give each information object a unique identifying name. This way relations of documents to those entities and, indirectly, to other documents are supposed to be detected. (Karger and Jones 2006, p. 82)

While techniques like Named Entity Recognition in electronic documents are already quite advanced, more sophisticated parsing of electronic document content is still problematic. One example are the current limitations of Natural Language Processing (NLP) techniques which are not sufficiently solved to work resource-efficiently with an acceptable margin of error or require an undue amount of manual effort to adjust and train the NLP algorithms (Castell et al. 2007, p. 14f). Another limitation to content based approaches stems from the fact that certain documents (e.g. multimedia files) contain content in proprietary formats which complicate the content extraction or make it altogether impossible.

2.2 Activity-based Relation Building

A completely different view of document relationship discovery is less focused on the documents themselves. Instead, the activities of a user around a document or section are deemed the critical information for discovering document relationships. Reading, editing, copying, pasting, sending email attachments or downloading files, indeed any action a user can take with a document, are used to discover key relationships. These approaches are mainly based on two assumptions: (a) the user switches between different activities are detectable, and (b) each activity is associated with a set of resources relevant to that activity. Resources can be documents, photographs, podcasts, email messages, web pages, RSS feeds, social bookmarks, chats and so on. Activity management systems help with task switching and resource (re-)discovery by providing a context for organizing and accessing related resources.

In TaskTracer (Dragunov et al. 2005) the user indicates when he begins a task and when the task is complete. TaskTracer then monitors documents and user activity to learn relevant folders and file locations, specific files manipulated, and a range of application settings relevant to the task. Once the user has done this the first time, on subsequent engagements with the task, TaskTracer will identify the active task and reset the state of relevant applications and documents. Another project, ActivityExplorer (Millen et al. 2005), takes a slightly different approach. In ActivityExplorer the user specifies the boundary of tasks by explicitly indicating the set of documents that are part of the task. In essence, this model has the user explicitly indicate how documents are related; there is no automatic relationship discovery. Explicit articulation of activity in ActivityExplorer, combined with tagging, has been combined in a search interface to exploit the user specified relationships. Relying on "tasks" as the principle means of document relationship discovery requires identifying the connections of one piece of information in one application to "task" related information in the same or other applications. Often a single task or workflow will require the use of multiple, differing applications, resulting in different interaction techniques and different representations to support the user.

Usage-Tracking has been deployed for a variety of purposes under a common rubric of looking for the user's intention. For instance, implicit feedback systems attempt to infer user intent based on observable behavior, such as which documents she does and does not select for viewing, and how long he views them (Oard and Kim 2001, p. 1f).

3 Our Approach

All of the above mentioned approaches operate in the domain of a single-user desktop environment. As mentioned in section 1 the potential redundancy of documents as well as the corresponding complexity to unravel the information-chaos rises exponentially when one looks at a multi-user environment. To the best of our knowledge no research project has applied Activity-based Relation Building to the domain of a multi-user environment as of yet.

Due to the described limitations of Content-based Relation Building our approach is set to avoid the ambiguities that arise when dealing with natural language texts or visual images in multimedia files in an automated way. Instead we are trying to utilize the capabilities of humans and computer systems where they fit best: The cognitive ability of humans to understand the deeper meaning of unstructured information and the analytical capabilities of computer systems to deal with large quantities of structured data. At the same time the whole process should be as transparent to the knowledge worker as possible to avoid additional workload on his part.

To this end we propose to integrate the different types of available document-related metadata, in particular static granular metadata (section 3.1) and dynamic contextual metadata (section 3.2) of electronic documents. As will be shown it becomes necessary to implement an initial seeding of external entity information to act as relationship-anchors (section 3.3).

3.1 Static Granular Document-Metadata

One special peculiarity of unstructured information in electronic documents is that those documents can be analyzed at different levels of granularity. Depending on the particular document type one chunk of information may be contained inside one sentence, span multiple paragraphs, or make up the whole document. Also multiple information chunks may rest inside one document instance. This becomes even truer when one considers an image or multimedia file with the different information chunks contained in sections such as foreground, background or specific subparts of the image.

Due to their multi-informational nature electronic documents are, in contrast to structured data, usually not automatically filed into single, one-dimensional taxonomies. Rather this classification falls to the human minds of knowledge workers who perform the cognitive task to classify these documents into one or more categories (Kwasnik 1989). To give an example a person might manually save an incoming email attachment into his local folder hierarchy. He may also store the attachment in multiple folders using symbolic links or creating redundant copies if the document contains information bits relating to different topics.

The granularity of many electronic documents is in itself additional meta-information that would not be present had the contained information, piece by piece, been extracted, structured and stored in a database system. A human user performed a cognitive effort when assembling multiple chunks of information into one document. This allows the assumption that those pieces of information are connected to each other semantically. A similar line of thought can be drawn regarding the filing and categorizing a human person does when organizing his personal archive, email management application or file system.

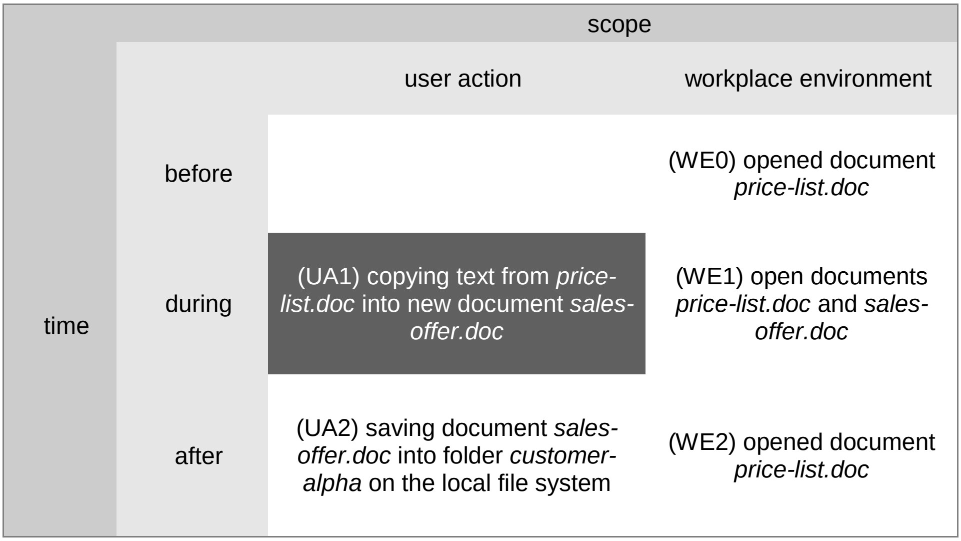
The, from a knowledge management point of view, at first sight complicating aspects of unstructured information in an electronic document may actually prove to be a possible point of leverage. They can, as in the saying the whole is more than the sum of its parts, provide additional meta-information about semantic relations the document may have regarding contained information or regarding semantic relations the document may have to other documents.

3.2 Dynamic Contextual Document-Metadata

In addition to the granular aspect of electronic documents further meta-information can be extracted from the context in which knowledge workers access and modify those documents (see section 2.2) (Barreau 1995, p. 329ff; Shen et al. 2005, p. 2f). In this paper the context of document access is understood as the sum of all actions taken by a user as well as the workplace environment shortly before, during, and after access of an electronic document.

The individual elements compromising such context can be separated into two dimensions: time and scope. The dimension of time is split into context elements which happened before, during, and after the document access. The dimension of scope differentiates among user actions and workplace environment. Of course the workplace environment and especially changes to it are always results of user actions. Therefore we restrict the definition of a user action to some input from the user which directly relates to a document, not to an application in itself. In contrast the workplace environment component in the context of a document access consists of all opened desktop and web applications as well as all content and documents visible in these applications. Additionally spontaneous communication by short email or instant messaging, which may not contain much meaningful information in itself, can act as glue by appearing in the same context as two or more other actions, linking them together.

To give a simple business scenario example: A user in the sales department copies some textual content from document *price-list.doc* into a new document *sales-offer.doc* and saves it in the local file system folder *customer-alpha*.



		scope	
		user action	workplace environment
time	before		(WE0) opened document *price-list.doc*
	during	(UA1) copying text from *price-list.doc* into new document *sales-offer.doc*	(WE1) open documents *price-list.doc* and *sales-offer.doc*
	after	(UA2) saving document *sales-offer.doc* into folder *customer-alpha* on the local file system	(WE2) opened document *price-list.doc*

Figure 1: Exemplary context of electronic document access

Figure 1 shows the context of the action UA1. The user action (in this case copying) allows the assumption (not certain knowledge) of an existing semantic relation between the documents *price-list.doc* and *sales-offer.doc*. The strength of this assumption as well as the type of semantic relation depend on heuristics based on typical usage patterns of document access. Additionally the metadata gained from observing user action UA1 might be influenced by the corresponding workplace environment WE1. So could for example a window-setup which shows both documents next to each other without overlap strengthen the assumption of an existing relationship. Part of the context of UA1 is the second user action UA2. This in turn can be used to establish a new (or strengthen an existing) semantic relation of *sales-offer.doc* and (to a lesser extend) *price-list.doc* to other documents contained in folder *customer-alpha*.

From a modelling perspective UA2 has its own context, which would then contain UA1 as an action that happened shortly before UA2.

3.3 External Business Entity Metadata

So far only one abstract entity-type was mentioned, document. Semantic relations among documents can be deduced by analyzing both the taxonomic placement in file hierarchies as well as the context of document access and modification. Both scenarios build upon the cognitive effort a knowledge worker put into those tasks.

For the resulting network of relations to have any use in the exemplary use case of E-Discovery a link to the business domain must be established. For this purpose we propose the insertion of external entity-types (e.g. product, process, role or person). These external entities, when linked to documents by semantic relations, can then be part of the context of a document access. The above described inferring of additional relations by analyzing the context of document access can in that case allow new assumptions to be made regarding semantic relations between external entities and documents. If we know document *price-list.doc* in the example shown in Figure 1 has a semantic relation to some specific product the assumption can be made that document *sales-offer.doc* also has a (however light or strong) relation to the same product.

The question is than how to initially link these external entities to documents. Fortunately certain relations are, in a usual enterprise environment, already present in machine readable ways. For example an Active Directory installation in combination with an Identity and Access Management (IAM) solution relates person records to product- or process-related roles. Also at least the centralized Document Management solutions often require users to tag uploaded documents or to enter metadata to place a documents main topic into a predefined taxonomy. Lastly collaborative project workspaces inherently infer a project-relation to the contained documents.

4 Aggregation and Analysis of Semantic Metadata

Section 3 mentioned the uncertainty which comes with every assumption made on semantic relations among entities. After all, without explicitly asking the specific user an automated system cannot know for sure if the exemplary documents *price-list.doc* and *sales-offer.doc* actually have any sort of semantic relation. The research hypothesis underlying this paper is that, similar to the approach of collaborative filtering, with a high number of data sets the amount of white noise of false positives can be reduced below an acceptable margin of error. The use case of this approach being a collaborative enterprise environment - with a high number of users and multiple occurrences of the same document in various repositories – has the secondary effect that the amount of both reliable[2] taxonomical as well as unreliable contextual and inferred (meta-) data might be immense. Before dynamic contextual information as described in section 3 can be utilized the basic methods of aggregating and analyzing semantic metadata in general need to be evaluated.

Section 4.1 describes the prototypical "ConSense" system to construct and aggregate basic semantic relations among documents in a multi-user scenario. Section 4.2 builds upon this prototype and shows how the immense amount of generated semantic (meta-) data can be handled. To evaluate the technical viability of the described approach the test case is limited to static document-based metadata. The processing of dynamic contextual metadata will have to be shown in a following paper.

4.1 Overview of the ConSense Prototype

To aggregate and analyze the assumptions made on semantic relations we implemented a distributed prototypical application "ConSense".

The business domain in this scenario is the ecosystem of knowledge workers in the enterprise accessing and modifying electronic documents as well as collaborating in teams and exchanging email messages.

Client-side plug-ins on the knowledge workers workstations extract static metadata from documents as well as detect and record the context of user activities related to document access and modification. Additionally sensors obtain the context of document-related user actions in the enterprises central document management applications and collaborative project workplace solutions. To ensure interoperability among different subsystems ontologies representing the set of concepts specific to the enterprise domain and the possible relationships among those concepts are used. The main ontology classes document, process, product, role and person can further be customized depending on industry-specific peculiarity. A good part of the Information-Lifecycle-specific abstract classes, as well as their possible relationships are imported from the FOAF and Dublin Core ontologies. The sensor plugins validate gathered context input against the common domain ontologies and persist it in the form of semantic networks in local Resource Description Framework (RDF) triple-stores. In the next step the sensor-plugins

[2] In this context the term reliable is used to differentiate first-order metadata from assumptions containing unreliable second-order (inferred) metadata.

add further statements based on the initially gathered static metadata by using heuristics based on typical usage patterns of document access. These statements contain assumption on the similarity or implicit correlation of either document-pairs or relations of documents to seeded business entities.

Document-related metadata with relevance beyond that of the local user (that is higher-level metadata relating to the domain-specific business entities) is then submitted to a central semantic information repository which is also implemented as an RDF triple-store. To reduce the amount of necessary network communication between a client and the central semantic store the transmission of RDF triples is selective. Even though both subsystems reference the same ontologies - meaning they have the same understanding of document and real-world entities and their possible interrelations – different subsets of the ontologies are in actual use. So will the central semantic store only consider aggregated and higher-level predicates which will actually be the target of subsequent discovery queries (regarding domain specific-questions) as input from the sensor plugins. This allows for baseline metadata to remain in the local RDF-stores on the clients while only higher-level metadata (the common ontology-subset) is transferred and made available to other users. Additionally the central semantic store – upon receiving sensor data - condenses this new data by merging statements with the same subject resource into the existing network. This is for example used to aggregate the reification-statements on the confidence of inferred interrelations from the client sensors.

The central semantic store, representing a virtualized view of assumptions on real-world relations among business-entities and documents, can in turn be queried by a user or knowledge manager to visualize and cluster relationship types according to his task-specific discovery needs. In the current implementation a *SPARQL/Update* endpoint is exposed as a query interface (a future prototype should allow for more sophisticated access and come with domain-specific preconfigured queries).

For example the litigation manager of the E-Discovery scenario could query for and detect documents having a relation to a product line being the subject in a legal low-product-quality complaint by a client. The query parameters could then be narrowed to documents having additional relations to the companies Quality Management process. Alternatively a rule- or heuristic based electronic service can directly access the semantic aggregation layer and query or manipulate the semantic network.

4.2 Coping with very large sets of inferred Metadata

To test the scalability of the used approach we ran an initial test of a prototypical client-side sensor-plugin on four test- clients. External entities were seeded in the RDF-encoded form of a corporate directory and a list of corporate products and services.

The sensor-plugins read and locally stored all static meta-attributes of documents[3] contained in the users "My Documents" folders. Extracted meta- attributes were then matched to a prepared set of RDF predicates referencing a subset of the FOAF and Dublin Core ontologies. The average number of documents on these test clients was in the range of 10,000. Including meta-attributes specific to Microsoft Office documents on average 50 RDF-triples per document were generated. This resulted on average in 500,000 triples containing reliable information. Comparing MD5 hashes of document content the overlap of reoccurring documents across clients lay at 8% of the total amount of detected documents. In the next step the sensor-plugins added further statements based on document name and location in the local file system. For example the titles of document- as well as folder-pairs were compared using the Smith-Waterman-Gotoh-Algorithm. When the similarity exceeded a minimum-threshold a statement containing the predicate "similarContent" connecting the two documents or folders in focus was added. In a similar way links to the business domain were created by matching properties of the initially seeded business entities to meta-attributes of documents and their containing folder taxonomies (Named Entity Recognition). Additionally these meta-statements containing unreliable assumptions were each referenced with a reification-statement of predicate-type "confidence" proportional to the numeric result of the respectively applied similarity-algorithm.

In the test setup this raised the total number of statements in the local RDF-store of each client by 400,000 to 900,000[4]. Seeing that an average RDF-statement in this test consisted of ~290 bytes (with dereferenced namespaces ~460 bytes) such an amount of RDF-triples is beyond an acceptable technical boundary both regarding transmission of data to a central semantic store as well as regarding the performance of further inferencing in the resulting semantic network.

Combining the aggregation-methods described in section 4.1 the number of RDF-statements persisting from the local RDF stores into the central semantic store could be reduced by 95%. In the test setup this resulted in a store of 160,000 statements of which 18,000 specifically described semantic relations of the 800 identified common documents on all four clients.

Decoupling the semantic repositories of clients and the central store has two further beneficial side effects: (1) It allows for system-resilience regarding outage or unavailability of individual sensors. (2) The sensors can filter data to be transmitted against predefined black- or white-lists and prevent too personal predicates ("hasRead", "visitedWebsite") from leaving the boundaries of the client desktop.

[3] In this test Microsoft Office files, multimedia files, pdf as well as postscript files, and text-files (including program source code) were considered.

[4] Further tests involving a dramatically higher document population (as could be expected in an enterprise document repository) of 200,000 documents resulted in the expected 10 million reliable statements and an additional 20 million unreliable assumptions using the same algorithms. The number of possible document-interrelations rose exponentially.

5 Conclusions and Outlook

The described approach is intended to utilize the cognitive acts knowledge workers perform during their everyday work, by analyzing the context of those acts and aggregating assumptions on document interrelations using semantic network techniques. In the test setup heuristics were used to condense the initially very large number of resulting semantic statements. This lowered the amount of data which needs to be transferred to a queriable central semantic store to an acceptable amount making further inference in the network technically viable.

In future research the underlying hypothesis of the feasibility to extract semantic relations from work-context will need further evaluation. Do humans in PC-based workplace environments really behave in such a way that the extracted assumptions on interrelations hold truth?

Also the mentioned heuristics to detect document-interrelations will have to be further tested and improved to reduce the number of false-positives. Here domain-specific heuristics might prove to be valuable. Lastly legislative aspects, especially privacy concerns of employees, have to be considered. Here a combination of white- and/or blacklisting named entities might be feasible, to specifically include business process relevant documents only or to exclude documents and communication of sensible parties from the context readings.

References

Adar E, Kargar D, Stein LA (1999) Haystack: per-user information environments. In Proc. CIKM'99, pp. 413-422.

Baron JR, Thompson P (2007) The search problem posed by large heterogeneous data sets in litigation: possible future approaches to research. ICAIL '07: Proceedings of the 11th international conference on Artificial intelligence and law, pp. 141-147.

Barreau D (1995) Context as a factor in personal information management systems. Journal of the American Society for Information Science 5(46), pp. 327-339.

Blumberg R, Atre S (2003) The problem with unstructured data. DM Review Magazine, http://www.dmreview.com/issues/20030201/6287-1.html, last accessed on 200909-05.

Castell P, Fernández M, Vallet D (2007) An adaptation of the vector-space model for ontology-based information retrieval. IEEE Transactions on Knowledge and Data Engineering, vol. 19, no. 2, pp. 261-272.

Dragunov AN, Dietterich TG, Johnsrude K, McLaughlin M., Li L, Herlocker, J (2005) TaskTracer: a desktop environment to support multi-tasking knowledge workers. In Proc. IUI'05. ACM Press.

Dumais ST, Cutrell E, Cadiz J, Jancke G, Sarin R, Robbins DC (2003) Stuff I've Seen: a system for personal information retrieval and re-use. In Proc. SIGIR'03, pp. 72-79.

Ferris Research (2008) Industry statistics. http://www.ferris.com/research-library/industry-statistics/. last accessed on 2009-07-20.

Gilbert MR, Shegda KM, Logan D, Chin K, Bell T, Latham L, Knox RE, Lundy J (2006) Key issues for enterprise content management. Gartner Research.

Guther M (2007) Duet - eine Software verknüpft Microsoft Office mit SAP-Geschäftsanwendungen. IM - Information Management & Consulting.

Karger DR, Jones W (2006). Data unification in personal information management. Commun. ACM 49 (1), pp. 77-82.

Kwasnik BH (1989) The influence of context on classification behavior. Doctoral thesis, Rutgers University.

Lyman A, Varian B (2003) How much information? Techreport, University of California,Berkeley.

Millen DR, Muller MJ, Geyer W, Wilcox E, Brownholtz B (2005) Understanding users and usage patterns: Patterns of media use in an activity-centric collaborative environment. In Proc. CHI 2005, ACM Press.

Murphy B (2006) Believe It - eDiscovery Technology Spending To Top $4.8 Billion By 2011. Forrester Research.

Oard DW, Kim J (2001) Modeling information content using observable behavior. In Proceedings of the 64 Annual Meeting of the American Society for Information Science and Technology.

Runyon B (2007) Data classification is a vital first step in information life cycle management. Gartner Research.

Shen X, Tan B, Zhai C (2005) UCAIR: Capturing and exploiting context for personalized search. Techreport, University of Illinois at Urbana Champaign.

Zadok E, Osborn J, Shater A, Wright C, Muniswamy-Reddy K, Nieh J (2004) Reducing storage management costs via informed user-based policies. 12th NASA Goddard, 21st IEEE Conference on Mass Storage Systems and Technologies (MSST 2004)