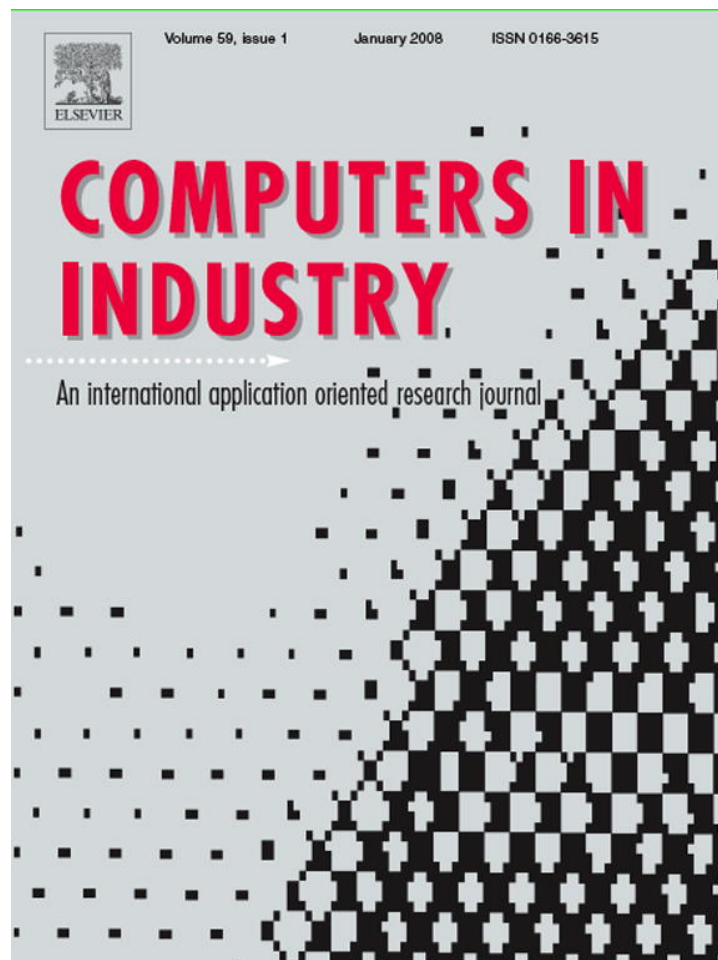


Provided for non-commercial research and education use.  
Not for reproduction, distribution or commercial use.



This article was published in an Elsevier journal. The attached copy is furnished to the author for non-commercial research and education use, including for instruction at the author's institution, sharing with colleagues and providing to institution administration.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

<http://www.elsevier.com/copyright>



## Review article

# A review of structured document retrieval (SDR) technology to improve information access performance in engineering document management

S. Liu <sup>a,\*</sup>, C.A. McMahon <sup>b</sup>, S.J. Culley <sup>b</sup><sup>a</sup> *University of Strathclyde, CAD Centre, DMEM Department, 75 Montrose Street, Glasgow G1 1XJ, UK*<sup>b</sup> *Department of Mechanical Engineering, University of Bath, Bath BA2 7AY, UK*

Received 9 November 2006; received in revised form 27 June 2007; accepted 24 August 2007

Available online 24 October 2007

**Abstract**

Information retrieval (IR) is a well-established research and development area. Document formats such as SGML (Standard Generalised Mark-up Language) and XML (eXtensible Mark-up Language) have become widely used in recent years. Traditional IR systems demonstrate limitations when dealing with such documents, which motivated the emergence of structured document retrieval (SDR) technology intending to overcome these limitations. This paper reviews the work carried out from the inception to the development and application of SDR in engineering document management. The key issues of SDR are discussed and the state of the art of SDR to improve information access performance has been surveyed. A comparison of selected papers is provided and possible future research directions identified. The paper concludes with the expectation that SDR will make a positive impact on the process of engineering document management from document construction to its delivery in the future, and undoubtedly provide better information retrieval performance in terms of both precision and functionality.

© 2007 Elsevier B.V. All rights reserved.

*Keywords:* Literature review; Structured document retrieval; Information access performance; Engineering document management**Contents**

1. Introduction	4
2. SDR key issues	4
2.1. Document structure study	4
2.1.1. Document structure definition	5
2.1.2. Document structure analysis	5
2.2. Document mark-up	6
2.2.1. Types of mark-up	6
2.2.2. Mark-up languages	6
2.2.3. Document mark-up strategies	7
2.3. Summary of document structure and mark-up from implementation view	8
3. SDR to improve information retrieval performance	8
3.1. SDR for information classification	8
3.2. SDR for information indexing	9
3.3. SDR for information querying and ranking	10
3.4. SDR for information presentation	10
3.5. Summary	10
4. Discussion and future research directions	10
4.1. Comparison and distribution of existing work	11

\* Corresponding author. Tel.: +44 141 548 2374; fax: +44 141 552 7986.

E-mail address: [shaofeng.liu@strath.ac.uk](mailto:shaofeng.liu@strath.ac.uk) (S. Liu).

4.2. Future research directions .....	12
5. Conclusion .....	13
Acknowledgements .....	13
References .....	13

## 1. Introduction

Engineering processes are highly creative and knowledge-intensive and comprise activities such as design, engineering analysis, manufacture and performance evaluation. As the complexity of these activities increases, so does the complexity of the information exchange and communication between engineers [1]. Several studies show that engineers spend as much as two-thirds of their time communicating in order to get input to their work and to output results from their work [2], and one-third of their time on searching for and accessing design information [3]. Many engineers seek information from plural sources such as documents, people and agents, depending on the stage of engineering process. For example, in the early design stage (such as conceptual design) when product features are vague, design engineers may focus more on seeking information from colleagues, communities and their own memories [4]. At the later stages such as detailed design and engineering analysis stages, engineers will rely more on formally recorded information such as reports, drawings, models and manuals [5,6].

It is believed that only 20% of formal information can be extracted from data warehouses comprising numeric data only, the other 80% of information is hidden in documents [7]. A similar observation has been made by Feldman [8], who claims that 80% of explicit knowledge in an enterprise can be found in documents. Therefore, document management has been recognised as a key topic in the information and knowledge management [9], which has been a well-established research field and has been successful in application for many areas. This paper provides a review on improving *document* access performance by integrating key approaches such as structure study and mark-up into an engineering domain. Readers who are interested in information seeking from *people* can refer to [10–12], study of how engineers' information seeking practices intertwine looking for informing documents with looking for informed people can be found in [2].

It is important to help engineers find the right information from documents at the right time in a way that most suits the engineers' profiles [13]. However, the information access performance of many document retrieval systems is limited by the technologies employed in the systems [14]. In traditional IR, systems make use of content information and focus on finding and delivering relevant whole documents (e.g. an entire paper or a Web page) in response to the user's needs. However, research has shown that the information needs of engineers are often not best met by the return of whole documents, but rather of the most relevant parts (of varying extent) of documents [15–17]. In particular, searching for information with traditional IR in very large documents is problematic, as is finding and

collating specific information from multiple document sources. To remedy this, structured document retrieval (SDR) technology has established itself in recent years as an active field of research and development which is distinguished from traditional IR in making use of both structural and content information, therefore allowing users to retrieve the most relevant *components* of documents, i.e. document content that is more focused on the users' information needs, for example, a section of a book instead of an entire book [18–20]. Moreover, appropriate representation of a structured document allows for the retrieval process to return *aggregated* components, for example a set of sections, or all sections of the documents that are relevant to a query, instead of delivering the whole document [21]. Therefore, by taking advantage of the structural information, SDR has two major gains: increased functionality and increased precision [22–24]. SDR is becoming increasingly important to engineering information and knowledge management as structured document formats such as SGML (Standard Generalised Mark-up Language) and XML (eXtensible Mark-up Language) are becoming widely used in enterprises for document publishing and distribution, using Web and other technologies [25].

This paper provides a review on various research activities carried out in the past decade involving the SDR in engineering information management. The core of the paper identifies two key SDR academic research areas: document structure study and mark-up. A comparison of selected papers is carried out based on exploitation of structural information and mark-up for improvement of information access performance in terms of classification, indexing, querying and ranking, and presentation. The final part provides a discussion on the gaps in the research and suggests future directions for SDR research in engineering document management.

## 2. SDR key issues

This section discusses two key issues of SDR that have a strong impact on information access performance. One is document structure study. The other is document mark-up.

### 2.1. Document structure study

Document structure is about how content objects are organized in a document such as books, scientific articles and technical manuals [26]. The structure of documents provides a new source of information in addition to document content that IR systems may exploit to improve their search effectiveness. This section starts with the definition of document structure, and then reviews the work on analysis of document structure.

### 2.1.1. Document structure definition

Establishing the understanding of document structure can be based on different viewpoints. Two crucial clarifications about document structure were introduced by Nunberg in his book *The Linguistics of Punctuation* [27]. First, *text structure* (for example, a sentence can be viewed orthographically as a portion of text starting with a capital letter and ending with a full stop) was distinguished from *syntactic structure* (for example, a sentence can also be defined by phrase-structure rules such as  $S = NP + VP$ , i.e. a sentence can be made up of a noun phrase and a verb phrase). Secondly, *abstract features* of text structure were distinguished from the *concrete* (or graphical) features by which they are expressed. Nunberg's "text-grammar" was recently extended by Power, Bouayad-Agha and Scott who proposed the use of *abstract document structure* as a separate descriptive level in the analysis and generation of written texts [28]. The purpose of this representation is to mediate between the message of a text (i.e. its discourse structure) and its physical presentation (i.e. its organisation into graphical constituents like sections, paragraphs, sentence, bulleted lists, figures and footnotes).

Document structures can be distinguished as visible or invisible. Visible structure refers to regular and systematic use of type, space and colour to expose a document's meaning to a human user, which is codified in grids or style manuals. Invisible structure refers to regular and systematic use of electronic markers to expose a document's meaning to a machine, which is codified in database tables and schemas, or in SGML/XML DTDs (document type definitions) and schemas [29].

Document structure can be internal (within a document) or external (across documents). Bratko and Filipic distinguished hyperlink structure from document internal structure [25], so did Hauff and Azzopardi [30]. Zhang and Radev addressed cross-document structures in their work [31].

In general, documents can be considered structured according to one or more structure types: linear order of words, sentences, paragraphs; hierarchy or logical structure of a book's chapters, sections; links (hyperlinks), cross-references, citations; temporal and spatial relationships in multimedia documents. Document structure analysis allows the structure types of a document to be identified.

### 2.1.2. Document structure analysis

The purpose of document structure analysis is to identify the document structure information of the source documents. There is a growing interest in document structure study because of the widespread use of structured documents (in contrast to flat documents, they have a logical structure and allow the incorporation of additional information through mark-ups, for example XML documents). From an IR point of view, a major advantage of structured documents over flat documents is the potential to access document elements or fragments. An earlier literature survey on document structure analysis algorithms can be found in [32]. Table 1 lists some recent important work on document structure studies. In Table 1, by application domains of image we mean image processing and image analysis for document scanning.

Generally, humans identify logical elements such as section headings or paragraphs from document images using geometric characteristics of the corresponding text areas. Additionally, by combining them into hierarchical structure elements such as sections, they recognise a logical structure of a document [33].

In [34], the authors propose a graph matching approach to label logical elements on a document page. A system developed by the authors can learn a model for a document class, use the model to label images through graph matching, and adaptively improve the model with error feedback. Lee, Choy and Cho present a syntactic approach for logical structure analysis of multi-page documents with hierarchical structures [35]. The work is considered as a valuable contribution to various fields such as digital library and office automation where transformation of large volumes of paper documents into electronic representations is needed. It is a step forward compared with its previous work that mostly targeted single-page documents [36,37]. Also aiming for logical structure analysis of document images, Ishitani proposed a method based on emergent computation [38]. This method has three principal advantages over other methods: adaptive system configuration for various and complex logical structures, robust document analysis tolerant of erroneous feature detection, and feedback of high level logical information to the low level physical process for accurate analysis. Experimental results obtained from 150 documents show that the method is adaptable, robust and effective for various document structures.

Table 1  
Some recent work on document structure analysis

Authors	Year	Structure types	Application domains	Performance or evaluation
Liang and Doermann [34]	2002	Logical structure	Image (single page document)	Tested on journal/proceeding articles, and the result shows promising accuracy
Lee et al. [35]	2003	Logical structure	Image (multi-page document)	Experiment with 372 images scanned from the IEEE transactions
Ishitani [38]	2005	Logical structure	Image	150 experimental documents, adaptable, robust, and effective
Wang et al. [39]	2005	Physical structure, Logical structure, Semantic structure	Text, automatic summarization	Flexible, systematic and extensible
Hauff and Azzopardi [30]	2005	Web structure	Web documents	Preliminary experiments, warrants further research
Zhang and Radev [31]	2005	Cross-document structure	Text classification	Helps improve performance of classifier

The above works mainly focus on logical structure of images of documents. Many more studies on document text identify document structure from other perspectives besides logical structure. For example, Wang, Wang and Gao defined a model for document structure analysis in three layers: physical structure layer, logical structure layer and semantic structure layer [39]. The model has been implemented on an Automatic Summarization System and good results have been achieved in the sense that the model is flexible, systematic and extensible.

The importance of structure of the World Wide Web has been widely recognised as more and more documents are available on Internet. Much research has been performed in investigating how links between Web pages can be exploited in an IR setting [40,41]. Hauff and Azzopardi explored the application of the Barabasi–Albert model to link structure analysis on a collection of Web documents within the language modelling framework [30]. The model utilizes the Web structure as described by a scale free network (a specific kind of complex network, in which some nodes act as “highly connected hubs” i.e. high degree, although most nodes are of low degree) and derives a document prior (document priors are probabilities or beliefs that the document is relevant to the query independent of any knowledge about the query) based on a Web document’s age and linkage. Preliminary experiments indicate the utility of the approach over other document link structure algorithms and suggest that it warrants further research.

Cross-document structure theory has emerged with the potential of improving various NLP (natural language processing) applications. An empirical study on cross-document structure relationships between sentence pairs extracted from topically related documents exploiting both labelled data and unlabelled data was reported [31]. The authors investigated a binary classifier for determining the existence of structural relationships and a full classifier using the full taxonomy of the relationships. The result shows that the exploitation of the cross-document structure helps improve the performance of learned classifiers.

The above work, with its emphasis on logical structure, physical structure, Web structure, etc., illustrates a successful case of understanding document composition from different viewpoints. However, the physical and logical views of document structure that predominate in the organisation of general documents are not likely to be sufficiently powerful to handle the wide range of objects found in engineering documentation. Current work on exploration of a more comprehensive view on document structures is being undertaken by the authors of this paper, based on decomposing documents from both content view and context view, from both physical view and logical view, and from technical view and media type view [42].

Once a document’s structure has been understood via its analysis from different viewpoints, the issue of how to take full advantage of the structural information in IR applications comes into the focus of many researchers, and mark-up has been recognised with great potential to facilitate the solution.

## 2.2. Document mark-up

It is important to make document structure explicit to ensure the accuracy of information retrieval. Document mark-up attempts to fulfil this role. This section describes the effort that has been made in the types of mark-up, mark-up languages, and finally mark-up strategies that support SDR in different circumstances.

### 2.2.1. Types of mark-up

Document mark-up was defined by Goldfarb as the process of adding codes to a document to identify the structure of a document or the format in which it is to appear [43].

Over last two decades, many mark-up methods have been developed and these can be classified into six basic categories [44,45]:

- (1) Punctuational mark-up—where word, phrase, and sentence boundaries are identified by spaces, commas, full stops, and other punctuation characters inserted into the text.
- (2) Presentational mark-up—where the visual form of the document is specified directly.
- (3) Procedural mark-up—in which presentational instructions (or commands) for some particular processing system are embedded in the text.
- (4) Descriptive mark-up—the author identifies the element types as tokens, as often found in applications of SGML and XML, which approach documents as structured objects containing semantically interpretable parts.
- (5) Referential mark-up—refers to entities external to the document and is replaced by those entities during processing.
- (6) Meta-mark-up—provides a facility for controlling the interpretation of mark-up and for extending the vocabulary of descriptive mark-up languages (e.g. macros).

### 2.2.2. Mark-up languages

A *mark-up language* is a formalised way of providing mark-ups [46]. There are dozens or maybe hundreds of mark-up languages in use today. Originally used for production within the publishing industry, mark-up languages have proliferated since the widespread adoption of XML and are being used to support information retrieval in many areas including science and technology [47,48]. This sub-section surveys the mark-up languages that are used to support SDR in three engineering circumstances: text retrieval, graphics retrieval, and Web services.

- (1) Mark-up languages for retrieval of text:

XML has been widely employed and implemented in engineering document content management for texts and tables. Most commonly, XML is used as a neutral format of information description and communication standard for engineering documents [56–58]. Further work in this stream, many modification versions of mark-up languages based on XML have been developed for different engineering scenarios. For example, DML (Dimensional

Mark-up Language) was discussed for closed-loop CAPP/CAM/CNC process chain based on STEP and STEP-NC [59]. To provide designers with a product data exchange standard and support product lifecycle management, Lee et al. proposed a Product Information Mark-up Language (PIML) [60]. A framework of XRML (eXtensible Rule Mark-up Language) developed by Kang and Lee allows the identification of rules on Web pages and extracts the identified rules automatically from Web pages to maintain consistency of text and table information [61]. In [62], a Motion Capture Mark-up Language (MCML) was proposed. The purpose of the MCML is not only to facilitate the conversion or integration of different formats, but also to allow for greater reusability of motion capture data, through the construction of a motion database storing the MCML documents.

## (2) Mark-up languages for retrieval of graphics:

Information retrieval from images and graphics has been an important issue for engineering information management. Traditionally, technical drawings, graphs and images were recognised by IR systems as a raster file and rendered to engineers as a whole picture. It used to be difficult for the systems to retrieve a specific part or text inside a technical drawing. The situation has changed since a series of mark-up languages were introduced over the years. Firstly, SVG [49], Scalable Vector Graphics, a novel XML-based language for describing two-dimensional graphics at the level of graphical objects rather than individual points, is now a W3C (World Wide Web Consortium) [50] standard. SVG is becoming increasingly popular due to its inherent advantages over raster image formats. SVG can work with three types of graphical objects: vectorial graphical shapes (such as lines and primitive shapes), text and images (equivalent to raster image retrieval). The creation of SVG was attributed from other two XML applications for encoding vector information. They are VML (Vector Mark-up Language) [51] and PGML (Precision Graphics Mark-up Language) [52]. VML defines a format for vector information together with additional mark-up to describe how that information may be displayed and edited, and is currently used by Google Maps for rendering vectors when running on IE 5.5 plus version. On the other hand, PGML is designed to meet precision needs of graphics artists besides simple vector graphics needs of casual users. Both VML and PGML were submitted to W3C as proposed standards in 1998, but failed, and then joined and improved upon to create SVG. X3D (eXtensible 3D) [53], the successor to VRML (Virtual Reality Modelling Language) [54], was introduced for real-time three-dimensional computer graphics and is now an ISO standard. X3D features many extensions to VRML such as Humanoid Animation, NURBS and GeoVRML, the ability to encode the scene using XML syntax as well as the Open Inventor-like syntax of VRML97, and enhanced application programmer interfaces (APIs). In recent years, SVG and X3D have been utilised in engineering domain such as design and manufacturing systems, especially for representation of

product models and provision with complex engineering data for applications such as CAD/CAPP/CAM [55].

## (3) Mark-up languages for Web services:

Web service was promoted as the biggest paradigm shift since the invention of Web itself [63]. Different types of Web services have emerged in engineering domain in particular for product design, business information systems and enterprise integration. Behind the success of all types of Web services, many XML-based standards such as SOAP (Simple Object Access Protocol), WSDL (Web Service Description Language) and XML-RPC (Remote Procedure Calls) are the key contributors. SOAP defines the XML-based message format that Web service-enabled applications use to communicate and interoperate with each other over the Web [64]. Based on SOAP, WSDL is a collection of metadata about XML-based services used to specify the procedures to discover functional and technical information about Web services over the Web [65]. XML-RPC is a remote procedure call protocol that uses XML to encode its calls and HTTP (Hypertext Transfer Protocol) as a transfer mechanism [66]. These XML-based standards enable applications to call functions from other applications, running on any hardware platform regardless of operating systems or programming languages, and have started finding their positions to play a central role in building and integrating enterprise applications. For instance, Umar discusses the application of SOAP, WSDL together with UDDI (Universal Description, Discovery and Integration) to enterprise boundaries business-to-business (B2B) trade and outsourcing through application service providers (ASPs) [67]. Comparatively, Bussler argues that the traditional Web service triplets SOAP, WSDL and UDDI allow enterprise application integration and B2B integration only possible at syntactic level, and therefore advocates using XML-RPC to enable semantic level integration [68].

### 2.2.3. Document mark-up strategies

Adding mark-ups to a document increases its value by making its content more accessible. Without mark-up, from a computer system's point of view, a document is just a long sequence of strings, and thus the operations that can be performed on such a document are rather limited. Once a document has been marked up, the system is able to exploit the implicit semantics of the mark-up tags, thus allowing for operations that are closer to the semantic level [69,70].

Document mark-up can be done at macro-level (dealing with the global visual and logical structure of a document), micro-level (used for marking single words or word groups), or symbol-level. Its practice is usually done manually, semi-automatically or automatically according to how much human effort is involved. Manual mark-up is the most popular, most accurate (with human interpretation) but most labour-intensive. Automatic mark-up reduces manual intervention and gives a more integral representation of the documents.

There have been limited publications concerning with automatic and hybrid mark-up. For example [71–73], all explored automatic mark-up with different methods and

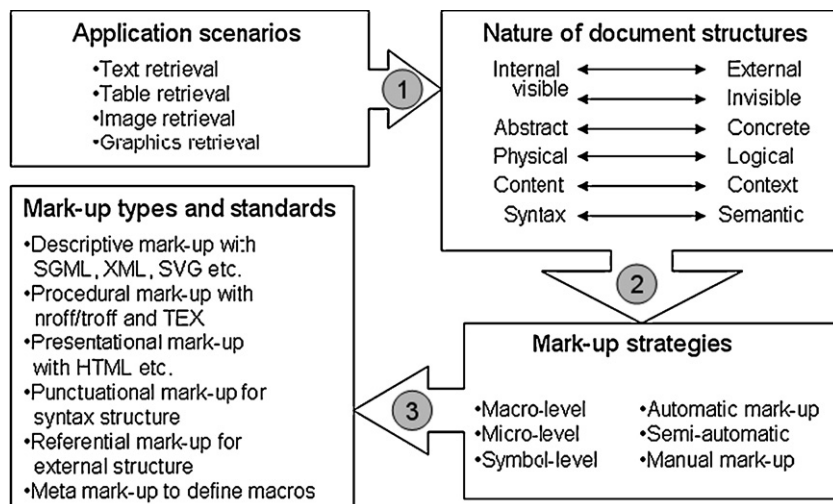


Fig. 1. Process to decide mark-up standards based on structures.

targeting on different types of documents. In [71], the authors reported on the design and implementation of a system which automates the process of capturing structured documents from the optically recognised form of printed materials to elements like words, sentences, title, authors, etc. but not for logical content elements. A novel system that can automatically mark-up text documents into XML is discussed in [72]. The system uses a self-organising map (SOM) algorithm and inductive learning algorithm. Experiments were carried out with business letters. The system can extract elements like address, date, salutation, paragraph and closing, etc. This system is adaptive in nature and learns from errors to improve mark-up accuracy. Cui discussed an automatic mark-up system which is based on machine learning methods and enhanced by machine learned domain rules and conventions [73]. The work is concentrated on taxonomic descriptions of plants (flora). Although machine learning systems are state-of-the-art, especially for simple tagging problems, knowledge based systems (mostly rule-based) have traditionally been the top performers in most information extraction benchmarks, and still retain some advantages. Feldman, Rosenfeld and Fresko propose a hybrid semantic tagging approach, which combines the power of knowledge-based and statistical machine learning [74]. The rules for the extraction grammar are written manually, while the probabilities are trained from an annotated corpus. The experiments show that the hybrid approach outperforms both purely statistical and purely knowledge-based systems. Vargas-Vera et al. present an annotation tool that provides both automated and semi-automated support for marking up Web pages with semantic contents [75].

### 2.3. Summary of document structure and mark-up from implementation view

This section has reviewed recent work on document structures and mark-ups. Different types of document structures and mark-ups are suitable for different application purposes. From implementation viewpoint, to decide what mark-up languages or standards for specific SDR applications is an

important matter, which undertakes three key steps, as illustrated in Fig. 1. First step is to determine document structures for the application scenarios, and then to determine mark-up strategies based on the nature of structures. Finally, to determine which mark-up standards and languages that can realise the mark-up strategies. Matrices can be developed to map the mark-up standards to mark-up strategies, to document structures, to application scenarios at each step. For details about the matrix development, please refer to authors' earlier publication [76].

### 3. SDR to improve information retrieval performance

The earlier sections of this paper have provided a broad review of the SDR approach and the key issues of structure study and mark-up technology. These establish the background for understanding of the current research in using SDR to improve information access performance. Until recently, SDR has been an academic domain, with most research work undertaken by computer scientists, even though general content-based document management approaches have been widely explored in engineering domain (for example [77–79]).

This section will survey the work on SDR both in academic research conducted by computer scientists and its engineering applications which involve using SDR technology to improve performance in terms of information classification, indexing, querying and ranking, and information presentation to users.

#### 3.1. SDR for information classification

Document classification has been used in IR in many different contexts such as document filtering, word sense disambiguation and classification against hierarchies as used in the Waypoint system [80]. Although most document categorisation research is limited to flat text representations, there exists a limited body of work that deals with structured documents. Generally speaking, the work in classification of structured documents makes use of two types of structure

information: internal structure information and hyperlink structure information.

Majority of existing work on classifying structured documents makes use of *hyperlink structure* information for HTML (Hypertext Mark-up Language) pages by employing flat text classifiers. For example, Furnkranz [81] and Glover et al. [82] consider using the textual context of links to a Web page, rather than text of Web page itself, as features for classification and show improvement in accuracy when classifying hypertext documents. Quek makes combined use of three classifiers operating respectively on the textual information, section titles and hyperlinks [83]. A similar approach is adopted in [84] with three classifiers respectively on HTML tags, metadata and textual information. Results show the improvement in the performance compared to a simple flat text classifier. Previous research by Ghani et al. [85] also shows that combining the use of hyperlink structure information and internal structure (tags and metadata) usually outperforms either representation alone. All this work relies on the HTML tag semantic and cannot adapt to more general structured document classification where for example the tag semantics is not known in advance.

To fulfil the categorisation of well structured documents such as in SGML and XML, more principled representations are needed to capture the *internal structure* information of the documents. In this area, researchers have been attempting to design new classifiers to adapt for structured documents. Yi and Sundarsan propose a vector representation for tree-like structured documents [86]. It makes use of the local word frequencies which depend on the element path from document root to the nodes and designs a Naive Bayes classifier to use on the resulting feature vector. The experiments were taken on two small document collections and with very basic structures. Denoyer and Gallinari use a splitting approach (a strategy to incorporate document structure, in which the flat text bags-of-words vector is split into a set of vectors, one for each component, each document is therefore represented by one such set of vectors. The splitting approach assumes that words are distributed differently in different document components and attempts to build a better classifier by modelling the distributions separately) in combination with a Naive Bayes text classifier [87]. This method was more extensively discussed in [88] using a SVM (support-vector-machine) classifier with a Fisher kernel on top of the original generative model. The results from experiments on a number of datasets with different types of structured documents are very encouraging. Piwowarski, Faure and Gallinari investigated using Bayesian networks for the task of ad-hoc retrieval of structured documents [89]. Recently, Bratko and Filipic examine three general approaches to exploiting structural information: tagging, splitting and stacking (a strategy to combine the predictions of individual per-component models from the splitting approach) [25]. In the stacking stage a meta-classifier makes the final prediction based on the results of different models obtained by splitting. The result shows that the relative success of these approaches is fairly stable across different base classifiers. In particular, the work finds that

stacking consistently performs best and often outperforms the flat text model by a wide margin.

Most previous systems for document management have relied on manual classification methods controlled by human experts. Due to the widespread use of information technologies for engineering and the increasing availability of electronic documents, manual classification is becoming unfeasible. More and more information classification systems developed for engineering document management are adopting advanced computer methods and artificial intelligence to achieve automation. For example, the Waypoint system, based on a faceted classification approach, is populated using an automatic constraint-based classifier [80]. Caldas, Soibelman and Han present a unique way, by using machine learning methods, to improve information organisation and access in inter-organisational systems based on automated classification of construction project documents according to their related project components [90].

### 3.2. SDR for information indexing

Index terms (or keywords, i.e. pre-selected terms which can be used to refer to the content of a document) play a distinct role in the IR. In the SDR the significance of index terms is strongly dependent on not only the frequency of the terms but also on the sections in which the terms occur. Furthermore, the terms are strongly dependent on the user needs. A limited amount of work in the literature has addressed these issues.

Bordogna and Pasi proposed a novel idea of defining an indexing mechanism which exploits a user's indication of their preferences on the documents' sections that they estimate to bear the most interesting information to create personalised and distinct representations of the same structured documents for distinct users [91]. This work has been further developed in [92], which presented a personalised indexing and retrieval approach for heterogeneous structured documents. The indexing mechanism comprised a dynamic component on top of a static component. The dynamic component is run at retrieval time during query evaluation. It computes the significance of the query terms in the document sections by taking into account the user needs expressed in the document structure-based and content-based conditions. The definition of this adaptive indexing mechanism constitutes a step forward towards the design of flexible IR systems allowing different levels of the adaptability to user needs.

Research in the engineering domain shows that high information retrieval performance of document management systems in terms of precision and recall can be achieved by appropriate indexing mechanisms regardless that the indexing process is operated manually or automatically. A concept space approach based on automatic indexing and concurrence analysis algorithms was introduced in [93] and was tested in the electrical engineering domain, generating significantly higher concept recall than the human-generated INSPEC computer engineering thesaurus. Comparatively, Yang, Wood and Cutkosky consider a manual indexing method to retrieve information captured from electronic design logbooks to meet

the evolutionary nature of design through the development of thesauri for design context [94]. They found that manual methods provide a higher level of IR performance, although with high overhead requirements. Further research on indexing design knowledge that is intuitive to an engineering designer contributed to the development of an ontology for engineering design [95].

### 3.3. SDR for information querying and ranking

To help users access and retrieve documents indexed with structural information, Egnor and Lord proposed a mechanism that uses reactive and automatic techniques to add structure to initially unstructured queries [96]. These techniques distinguish this work from database systems that presuppose proactively structured queries expressed in a formal query language such as SQL (Structured Query Language) or XML-QL (XML Query Language). A query mechanism that applies database style query evaluation in response to IR style keyword-based queries for retrieving relevant answers from a logically structured document is proposed by Pradhan and Tanaka [97]. The semantics of keyword-based queries are first defined. Then an algebra that is capable of computing every relevant portion of a document (which can be considered answers to a set of arbitrary keywords) is discussed. The ordering and structural relationships among the components are preserved in the answer. The paper also introduces several practically useful filters that save users from having to deal with an overwhelming number of answers.

An SDR system should score a document not only by term frequencies but also by examining its structure. Kim and Zhang propose a ranking scheme for HTML documents where ranking is not based on the existence of a term in a document, but the existence of a term in specified tags in the document [98,99]. The rank for a given document is the product of two scores: one for the term and one for the tag. The limitation of this scheme is that terms outside the chosen tags cannot be found. For terms inside the chosen tags, there is a risk of indiscrimination, for example, a document containing a search term twice in tag <title> and once in tag <paragraph> can have the same score as that twice in <paragraph> and once in <title>. This issue has been addressed by Trotman who presented a method of indexing and searching structured data allowing structure weighting, in which document structures are identified first and then term frequency is replaced by a structured weighted term frequency [24]. Experiments with weighted ranking using vector and probabilistic retrieval show significant improvements over un-weighted retrieval. A model proposed in [92] can flexibly query documents with heterogeneous structured documents. The term “heterogeneous structured documents” means the documents having distinct structures, i.e. differing in some or even all their constituent components or sections.

### 3.4. SDR for information presentation

It is well understood that information seeking is a vague and imprecise process, especially in SDR because structured documents are often long and complex. A good presentation

of information with well-designed user interfaces is found to be particularly effective in this case [100]. It has also been recognised that the best approach to querying structured documents is to let the user specify in the most natural way both the content and the structural requirements of the desired documents [101]. This issue has been addressed by [102] and further explored in [26]. In this latter work, the authors present a new GUI, with explanatory and selective feedback capabilities, for structured document retrieval specifically designed for hierarchically structured documents. A user-task oriented evaluation shows that the proposed interface provides the user with an intuitive and powerful set of tools for structured document searching, retrieved list navigation, and search refinement. Work reported in [103] tackles the problem of bootstrapping engineering document recognition systems. A user-friendly interface is presented aiming to acquire knowledge concerning the graphical appearance of objects and to learn the best approach to recognise the learned objects.

Relevance feedback is a desirable feature of modern information retrieval system. It covers a range of techniques intending to improve a user's query and facilitate retrieval of information relevant to a user's information need. Ruthven and Lalmas considered specific interfaces to relevance feedback systems and characteristics of searchers that can affect the use and success of relevance feedback systems [104]. Further in [105], the authors present five user experiments on incorporating behavioural information into the relevance feedback process. Models for representation and retrieval of heterogeneous structured documents can be found in [106,107].

### 3.5. Summary

Different SDR mechanisms have been reviewed in this section with respect to information classification, indexing, querying and ranking, and presentation. From information user's point of view, information (i.e. document fragment) retrieval undergoes key process steps such as forming information need in a form that can be understood by the SDR retrieval mechanisms, translating the information need into an appropriate query format to retrieve desired results. From information implementer's point of view, important process steps include describing documents in a form that allows the retrieval mechanism to identify the potentially relevant document fragments quickly and with precision based on the document structure analysis and mark-up. Here the key is to match information provider's implementation mechanisms with the user's information need, as shown in Fig. 2. For information implementers to start the SDR process, preparation is needed to mark-up the documents so that document fragments can be extracted ready for classification.

## 4. Discussion and future research directions

The authors have classified the numerous research interests of SDR to improve information access performance and its application to engineering document management referred in the paper into several areas. Fig. 3 shows a classification, which

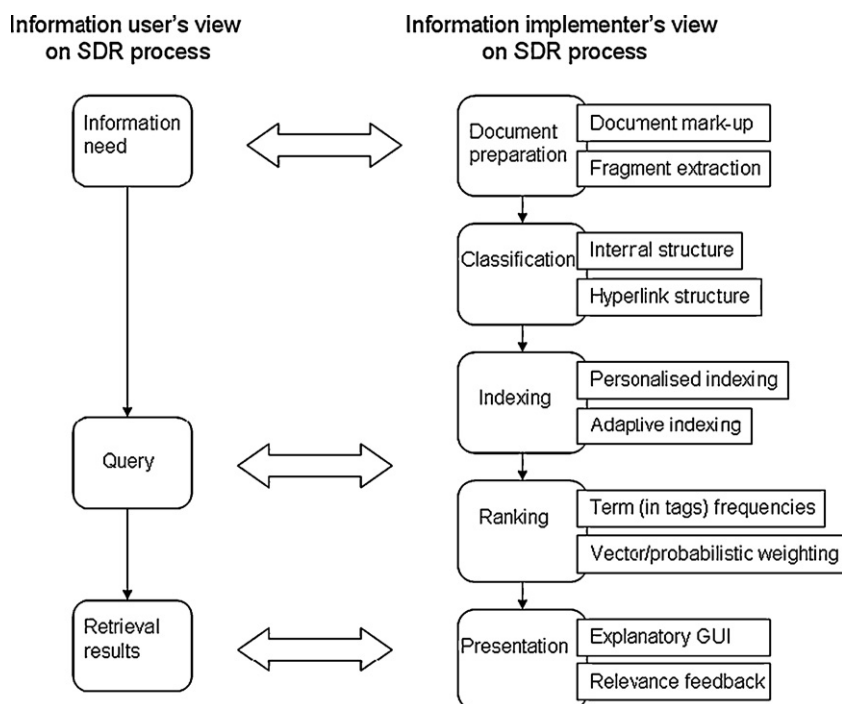


Fig. 2. SDR processes from both user's and implementer's view.

is used in this section to discuss the distribution of existing work, to compare interests between computer scientists and engineering researchers, and to identify possible future research directions.

#### 4.1. Comparison and distribution of existing work

The first comparison of the research is carried out based on the denominators (structures and mark-ups) that result in performance improvement in document retrieval and the specific application domains. For example:

- What structural information (internal or hyperlink structure) is used to support SDR process?
- What type of mark-ups are employed?
- What aspects of the performance are improved by employing the key denominators?

- What application areas (including engineering) have been targeted by existing work?

The key denominators against performance of previously discussed work are illustrated and compared in Table 2. Key denominators are primarily classified based on the type of structures and mark-ups. The structure types can be physical, logical, semantic, ontology, hyperlink, meta-data, etc., and mark-ups can be generic (SGML, XML) or specific. The performance aspects have been classified into four categories: information classification, indexing, querying and ranking, and presentation.

A wide diversity of structures has been considered by researchers. Physical and logical structures have been the most popular. At a higher level of importance, such structures as hyperlink, ontology and meta-data are also included. Accordingly, the mark-ups employed also show a wide range of

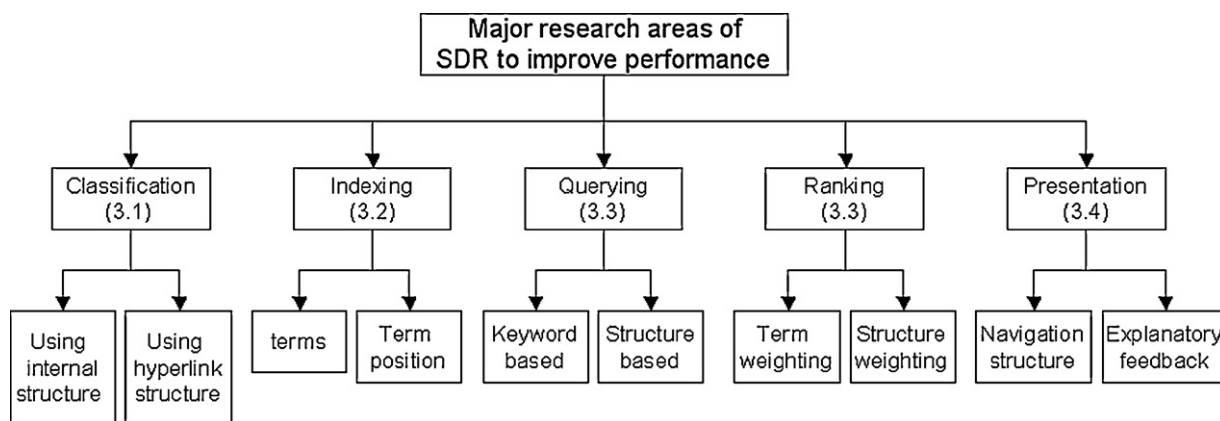


Fig. 3. Classification of SDR areas and applications (corresponding section numbers are in brackets).

Table 2  
Comparison of existing work based on denominators against performance

Authors	Structures	Mark-up	Performance aspects	Application domains
Law et al. [108]	Logical structure, hyperlink structure	HTML	Information presentation	Engineering document version control
Kim and Compton [109]	Hyperlink structure	HTML	Information classification, information presentation	Web personal pages
Shin [110]	Meta-data structure	XML	Information indexing	PVR environment
Kotinurmi et al. [111]	Meta-data structure	XML	information classification, information presentation	Product development
Gardoni et al. [22]	Meta-data structure, ontology structure	XML, RDF	Information indexing, information presentation	Aeronautical defence and space
Jokinen et al. [112]	Meta-data structure	XML	Information classification, information presentation	Product development
Thirunarayan et al. [113]	Ontology structure	XML	Information classification	Materials and process specifications
Wang et al. [39]	Physical structure, logical structure, semantic structure	HTML	Information ranking	Document summarisation
Brand et al. [114]	Ontology structure	OWL	Information classification	Design process
Wong et al. [115]	Ontology structure	RDF, OWL	Information indexing	Engineering design
Liu et al. [42]	Physical structure, logical structure, hyperlink structure, meta-data structure	XML, SVG, HTML	Information classification, information indexing, information querying, information presentation	Engineering design

approaches. Even though XML and HTML show clear dominance, RDF (Resource Description Framework) and OWL (Web Ontology Language) also have gained popularity, especially in the research related to meta-data and ontology. SVG has obtained its unique place with respect to graphics mark-up. In addition to the diversity of structure, Table 2 also shows that the researches using the same set of structures and mark-up for the same application domain were very scarce.

The second comparison is carried out between the existing work undertaken by computer scientists and engineering researchers, and the distribution among different aspects of information access performance, as demonstrated in Fig. 4.

This figure compares the work conducted by computer scientists and in engineering application in all four aspects of SDR to improve information access performance. From this figure, we can see two clear trends. First, it shows that the

research has been dominated by computer academia. The authors selected over 90 publications, only 16 out of them are concerned with SDR application in engineering, which counts about 17% of the collection. In fact, publications concerned with SDR application to engineering in all four aspects of information access performance, i.e. classification, indexing, querying and ranking, and presentation, are all numbered below 6. Secondly, in the research of computer scientists, the biggest interests are on information classification and presentation (user interface and relevance feedback). Comparatively, scientists show less interest in the information indexing, querying and ranking with SDR. This pattern is also reflected by the engineering applications.

4.2. Future research directions

With the development of structured textual and multimedia documents, and with the increasing importance of structured document formats like XML, the nature of documents is changing. There is an urgent need for realistic and effective strategies to establish SDR systems in engineering document management to face the changes and to take advantage of the opportunities those changes present. There has been active research to explore the key issues and methodologies to improve information access performance in engineering document management, many research questions are still open for discussion. The future research directions include (but not limited to):

- (1). As document structure has been fundamental to all the aspects of performance improvement, structure study should be expected to remain as an active research topic in the future. Existing work on structure study has been focusing on logical structure, physical structure and Web

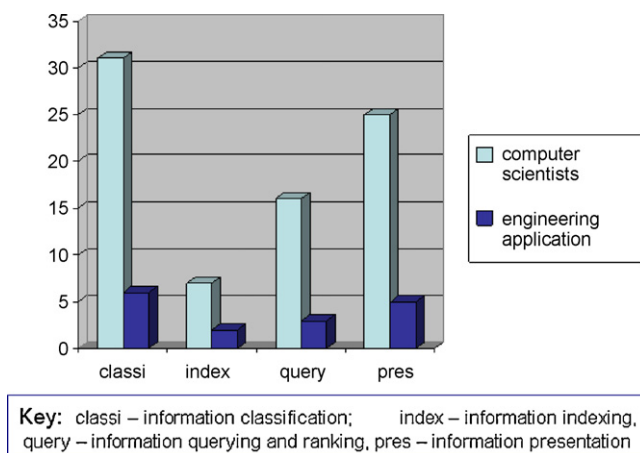


Fig. 4. Comparison between research by computer scientists and engineering researchers.

structure. Study of document structure from new perspectives or from more comprehensive perspectives should be in the heart of future work [116]. Considering that the dependencies between elements could be complicated for a large structured engineering document, further exploration of learning dependencies, especially the semantic analysis of the relationships between document elements, can still be a challenge for future work.

- (2). Structured documents usually have a much richer representation than flat ones, how to capture and exploit the structural information, especially for a complex and highly structured engineering documents such as technical reports or industrial process plans, can be difficult. Document templates have been widely used in industry document engineering practice, further work to explore the utilisation of mechanisms like document templates to assist in identifying, capturing and exploitation of structural information for engineering document retrieval is highly desirable. Software tools like Microsoft InfoPath allow users to define document templates compatible with XML schemas or DTDs, to draft documents using a graphical interface and to save documents in XML format. Development of similar software tools that can provide easy-to-learn mechanisms like templates for engineers to capture and exploit document structures will undoubtedly facilitate SDR technology to be further accepted in industry.
- (3). It is also important to work on best entry points [117] into a hierarchical structure in the SDR, given the aggregated evidence in individual objects. Research undertaken by [118] explored engineers' preference for elements in the document structure to show document content when accessing design documents. The results show that elements like headings, figures and tables are among engineers' preferred elements. Further work on the investigation of mechanisms to implement those preference elements as best entry points would provide SDR with tremendous engineer significance, because in this way presentation results to engineers will be focused by displaying only selected document components in the interface, rather than all relevant document components, thus reducing engineers' time and disorientation caused by lack of domain knowledge.
- (4). Companies have huge number of legacy documents that may not be constructed in a standard structure. Retrieval of information pieces (document fragments) from those legacy documents is not trivial work. To enforce a DTD or schema that represents document structure applicable to specific engineering scenarios into the legacy documents would be highly beneficial for engineering information management. Marking up documents without a DTD or schema for SDR is currently undertaken through artificial intelligence (for example, machine learning). As a next step forward, work should explore how automatic mark-up approaches take account of DTDs or schemas that fully represent engineering scenarios.
- (5). There is a trend from finding document components for specific engineering questions to creation of new answers.

After data warehousing and data mining have achieved success in engineering such as in manufacturing enterprises [119,120], many researchers now advocate it is time to conduct research works on document warehousing to capture business intelligence [7]. Unlike traditional document management, document warehouses include extensive semantic information about documents, cross-document feature relations, and document grouping or clustering to provide more accurate and more efficient access to text-oriented business intelligence.

## 5. Conclusion

The introduction of SDR to engineering information management has been a viable option for retrieving documents at component level and with precision. This is warranted by exploitation of structural information and mark-up technologies in the process of information classification, indexing, querying and ranking, and information presentation to users. With continuous advance of SDR fundamentals, the viability of SDR application to engineering document management can be considerably extended.

This paper has reviewed the background, the current state of the art of SDR approaches, key issues and its positive impact on the information retrieval performance. This review has been undertaken to identify the aspects that are applicable to SDR within engineering document management. Selected collections of publications have been analysed to establish the key requirements that determine information access performance, and to establish an initial view on how SDR has started changing the process of engineering document management, not only in the way information is retrieved, but also from the way a document is constructed to the way it is delivered, although exploration is still ongoing in this aspect. It can be expected that the key strategies of whole engineering document management processes will be substantially affected along with the SDR advance in the future.

## Acknowledgements

The research reported in this paper was funded by the UK Engineering and Physical Sciences Research Council (EPSRC) under grant number EP/C534220/1 for the project of "Immortal Information and Through-Life Knowledge Management: Strategies and Tools for the Emerging Product-Service Paradigm (also known as KIM project)".

## References

- [1] S. Liu, R.I.M. Young, Utilising information and knowledge models to support global manufacturing co-ordination decisions, *International Journal of Computer Integrated Manufacturing* 17 (6) (2004) 479–492.
- [2] M. Hertzum, A.M. Pejtersen, The information-seeking practices of engineers: searching for documents as well as for people, *Information Processing and Management* 36 (5) (2000) 761–778.
- [3] C.A. McMahon, A. Lowe, S.J. Culley, Characterising the requirements of engineering information systems, *International Journal of Information Management* 24 (2004) 401–422.

- [4] C.A. McMahon, A. Lowe, S.J. Culley, Knowledge management in engineering design: personalization and codification, *Journal of Engineering Design* 15 (4) (2004) 307–325.
- [5] A.G. Gunendran, R.I.M. Young, An information and knowledge framework for multi-perspective design and manufacture, *International Journal of Computer Integrated Manufacturing* 19 (4) (2006) 326–338.
- [6] C.A. Costa, R.I.M. Young, Product range models supporting design knowledge reuse., *Proceedings of the Institute of Mechanical Engineers, Part B: Journal of Engineering Manufacture* 215 (3) (2001) 323–337.
- [7] F.S.C. Tseng, Design of a multi-dimensional query expression for document warehouses, *Information Sciences* 174 (1–2) (2005) 55–79.
- [8] R. Feldman, Text mining: theory and practice, in: *Proceedings of the Fourth World Congress on Expert Systems—Application of Advanced Information Technologies*, ITESM Mexico City Campus, 1998.
- [9] D.A. Guerra-Zubiaga, A manufacturing model to enable knowledge maintenance in decision support systems, PhD Thesis, Wolfson School of Mechanical and Manufacturing Engineering, Loughborough University, UK, 2004.
- [10] J.Y. Lee, S.H. Hahn, J.S. Yoon, A study on the knowledge community for creating and sharing implicit knowledge, *On the Convergence of Bio-Information-, Environmental-, Energy-, Space- and Nano-Technologies, Parts 1 and 2: Key Engineering Materials* 277–279 (2005) 331–336.
- [11] M. Hertzum, H.H.K. Andersen, V. Andersen, C.B. Hansen, Trust in information sources: seeking information from people, documents and virtual agents, *Interacting with Computers* 14 (5) (2002) 575–599.
- [12] S. Robertson, Requirements trawling: techniques for discovering, *International Journal of Human-Computer Studies* 55 (4) (2001) 405–421.
- [13] S. Ahmed, K.M. Wallace, Identifying and supporting the knowledge needs of novice designers within the aerospace industry, *Journal of Engineering Design* 15 (5) (2004) 475–492.
- [14] H.K. Lin, J.A. Harding, P.C. Teo, An inter-enterprise semantic Web system to support information autonomy and conflict moderation, *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture* 219 (12) (2005) 903–911.
- [15] A. Lowe, C.A. McMahon, T. Shan, S.J. Culley, A method for the study of information use profiles for design engineers, *ASME Design Theory and Methodology*, Las Vegas, DTM-8753, 1999.
- [16] M. Aurisicchio, P.M. Langdon, S. Ahmed, K.M. Wallace, Investigating knowledge searches in aerospace design, in: *ICED03*, in: *Proceedings of the 14th International Conference on Engineering Design*, Stockholm, Sweden, (2003), pp. 293–294.
- [17] R. Fidel, M. Green, The many faces of accessibility: engineer's perception of information sources, *Information Processing and Management* 40 (3) (2004) 563–581.
- [18] T. Roelleke, M. Lalmas, G. Kazai, I. Ruthven, S. Quicker, The accessibility dimension for structured document retrieval, *Advances in Information Retrieval Lecture Notes in Computer Science* 2291 (2002) 284–302.
- [19] T. Shimizu, M. Yoshikawa, Full text and structural XML indexing on B-tree. Database and expert systems applications, *Proceedings Lecture Notes in Computer Science* 3583 (2005) 451–460.
- [20] J. Reid, M. Lalmas, K. Finesilver, M. Hertzum, Best entry points for structured document retrieval—Part I. Characteristics, *Information Processing and Management* 42 (2006) 74–88.
- [21] G. Kazai, M. Lalmas, J. Reid, Construction of a test collection for the focussed retrieval of structured documents, *Lecture Notes in Computer Science* 2633 (2003) 88–103.
- [22] M. Gardoni, C. Frank, F. Vernadat, Knowledge capitalisation based on textual and graphical semi-structured and non-structured information: case study in an industrial research centre at EADS, *Computers in Industry* 56 (2005) 55–69.
- [23] A. Trotman, Searching structured documents, *Information Processing and Management* 40 (4) (2004) 619–632.
- [24] A. Trotman, Choosing document structure weights, *Information Processing and Management* 41 (2) (2005) 243–264.
- [25] A. Bratko, B. Filipic, Exploiting structural information for semi-structured document categorization, *Information Processing and Management* 42 (3) (2006) 679–694.
- [26] F. Crestani, J. Vegas, P. Fuente, A graphical user interface for the retrieval of hierarchically structured documents, *Information Processing and Management* 40 (2) (2004) 269–289.
- [27] G. Nunberg, The Linguistics of Punctuation No. 18 in *CSIL Lecture Notes*, CSIL Publications, Stanford, CA, 1990.
- [28] R. Power, N. Bouayad-Agha, D. Scott, Document structure, *Computational Linguistics* 29 (2) (2003) 211–260.
- [29] Text Matters. Structure in document: an introduction, 2006. [http://www.textmatters.com/resources/pdfs/struct\\_docs.pdf](http://www.textmatters.com/resources/pdfs/struct_docs.pdf).
- [30] C. Hauff, L. Azzopardi, Age dependent document priors in link structure analysis, *Advances in Information Retrieval Lecture Notes in Computer Science* 3408 (2005) 552–554.
- [31] Z. Zhang, D. Radev, Combining labelled and unlabelled data for learning cross-document structural relationships. *Natural Language Processing—IJCNLP, Lecture Notes in Computer Science* 3248 (2005) 32–41.
- [32] S. Mao, A. Rosenfeld, T. Kanungo, Document structure analysis algorithms: a literature survey, 2002. <http://www.archive.nlm.nih.gov/pubs/mao/mao03.pdf>.
- [33] K.M. Summers, Toward a taxonomy of logical document structures, in: *Proceedings of DAGS'95*, 1995, pp. 124–133.
- [34] J. Liang, D. Doermann, Logic labelling of document images using layout graph matching with adaptive learning. *Document Analysis System V, Proceedings Lecture Notes in Computer Science* 2423 (2002) 224–235.
- [35] K.H. Lee, Y.C. Choy, S.B. Cho, Logical structure analysis and generation for structured documents: a syntactic approach, *IEEE Transactions on Knowledge and Data Engineering* 15 (5) (2003) 1277–1294.
- [36] D. Niyogi, S.N. Srihari, An integrated approach to document decomposition and structural analysis, *International Journal of Imaging Systems and Technology* 7 (1996) 330–342.
- [37] C. Lin, Y. Niwa, S. Narita, Logical structure analysis of book document image using contents information, in: *Proceeding of the 4th International Conference of Document Analysis and Recognition*, vol. 2, 1997, pp.1048–1051.
- [38] Y. Ishitani, Logical structure analysis of document images based on emergent computation, *IEICE Transactions on Information and Systems* E88D 8 (2005) 1831–1842.
- [39] Z.Q. Wang, Y.C. Wang, K. Gao, A new model of document structure analysis. *Fuzzy systems and knowledge discovery, Proceedings of Lecture Notes in Artificial Intelligence* 3614 (2005) 658–666.
- [40] R. Baeza-Yates, F. Saint-Jean, C. Castillo, Web dynamics, age and page quality, in: *Proceedings of SPIRE Lecture Notes in Computer Science*, 2002.
- [41] W. Kraaij, J.Y. Nie, M. Simard, Embedding web-based statistical translation models in cross-language information retrieval, *Computational Linguistics* 29 (3) (2003) 381–419.
- [42] S. Liu, C.A. McMahon, M.J. Darlington, S.J. Culley, P.J. Wild, An approach for document fragment retrieval and its formatting issue in engineering information management, *Lecture Notes in Computer Science* 3981 (2006) 279–287.
- [43] C.F. Goldfarb, GML: a generalized approach to document mark-up, in: *Proceedings of the ACM SIGPLAN SIGOA Symposium on Text Manipulation*, Portland, Oregon, 1981.
- [44] R. Khare, A. Rifkin, The origin of document species, *Computer Networks and ISDN Systems* 30 (1–7) (1998) 389–397.
- [45] J.H. Coombs, A.H. Renear, S.J. DeRose, Review of mark-up systems and the future of Scholarly text processing, 1996. <http://www.ei.cs.vt.edu/~cs5604/f96/art-summs/COOM.txt>.
- [46] P. Johnston, What you have always wanted to know that about SGML, HTML and XML but were afraid to ask: why mark-up matters, in: *Society of Archivists' Diploma in Archive Administration seminar*, Edinburgh, November 1998.
- [47] Z. Lu, A survey of XML applications on science and technology, *International Journal of Software Engineering and Knowledge Engineering* 15 (1) (2005) 1–33.
- [48] R. Kay, Mark-up languages, *Computer World*, August 8, 2005.
- [49] Y. Wang, P.N. Ajoku, J.C. Brustoloni, B.Q. Nnaji, Intellectual property protection in collaborative design through lean information modelling

- and sharing, *Journal of Computing and Information Science in Engineering* 6 (2) (2006) 149–159.
- [50] N. Anwar, W. Kanok-Nukulchai, D.N. Batanov, Component-based, information oriented 3D structural engineering applications, *Journal of Computing in Civil Engineering* 19 (1) (2005) 45–57.
- [51] N. Swindells, Communication materials information: product data technology for materials, *International Materials Review* 47 (1) (2002) 31–46.
- [52] C. Brecher, M. Vitr, J. Wolf, Closed-loop CAPP/CAM/CNC process chain based on STEP and STEP-NC inspection tasks, *International Journal of Computer Integrated Manufacturing* 19 (6) (2006) 570–580.
- [53] C.K.M. Lee, G.T.S. Ho, H.C.W. Lau, K.M. Yu, A dynamic information schema for supporting product lifecycle management, *Expert Systems with Applications* 31 (1) (2006) 30–40.
- [54] J. Kang, J.K. Lee, Rule identification from Web pages by the XRML approach, *Decision Support Systems* 41 (1) (2005) 205–227.
- [55] H.S. Chung, Y. Lee, MCML: motion capture mark-up language for integration of heterogeneous motion capture data, *Computer Standards and Interfaces* 26 (2) (2004) 113–130.
- [56] J.D. Eisenberg, *SVG Essentials*, O'Reilly: Beijing, Cambridge, Farnham, Köln, Paris, Sebastopol, Taipei, Tokyo, 2002.
- [57] W3C, World Wide Web Consortium, [www.w3c.org](http://www.w3c.org).
- [58] VML, Vector Mark-up Language, [www.w3.org/TR/NOTE-VML.html](http://www.w3.org/TR/NOTE-VML.html).
- [59] PGML, Precision Graphics Mark-up Language, [www.w3.org/TR/1998/NOTE-PGML-19980410](http://www.w3.org/TR/1998/NOTE-PGML-19980410).
- [60] X3D, eXtensible 3D, [www.web3d.org](http://www.web3d.org).
- [61] C.Y. Kim, N. Kim, Y. Kim, S.H. Kang, P. O'Grady, Distributed concurrent engineering: Internet-based interactive 3-D dynamic browsing and mark-up of STEP data, *Concurrent Engineering—Research and Applications* 6 (1) (1998) 53–70.
- [62] W.D. Li, A web-based service for distributed process planning optimisation, *Computers in Industry* 56 (3) (2005) 272–288.
- [63] D.S. Linthicum, *Next Generation Application Integration: from Simple Information to Web Services*, Addison-Wesley, 2003.
- [64] SOAP, Simple Object Access Protocol, <http://www.w3.org/2000/xml/Group/>.
- [65] WSDL, Web Service Description Language, <http://www.w3.org/2002/ws/desc/>.
- [66] XML-RPC, XML—Remote Procedure Calls, <http://www.xmlrpc.com/>.
- [67] A. Umar, The emerging role of the Web for enterprise applications and ASPs, *Proceedings of the IEEE* 92 (9) (2004) 1420–1438.
- [68] C. Bussler, Semantic Web services—fundamentals and advanced topics, *Lecture Notes in Computer Science* 3263 (2004) 1–8.
- [69] M. Abolhassani, N. Fuhr, N. Govert, Information extraction and automatic mark-up for XML documents. *Intelligent Search on XML Data*, *Lecture Notes in Computer Science* 2818 (2003) 159–174.
- [70] K. Taghva, R. Beckley, J. Cooms, The effects of OCR on the extraction of private information. Document analysis systems VII, *Proceedings Lecture Notes in Computer Science* 3872 (2006) 348–357.
- [71] K. Taghva, A. Condit, J. Borsack, Autotag: a tool for creating structured document collections from printed materials. <http://www.isri.unlv.edu/publications/isripub/Taghva98b.pdf>.
- [72] S. Akhtar, R.G. Reilly, J. Dunnion, Auto-tagging of text documents into XML. Text, speech and dialogue, *Proceedings Lecture Notes in Artificial Intelligence* 2807 (2003) 20–26.
- [73] H. Cui, MARTT: a general approach to automatic mark-up of taxonomic descriptions with XML. [http://www.caais-acs.ca/proceedings/2005/cui\\_2005.pdf](http://www.caais-acs.ca/proceedings/2005/cui_2005.pdf).
- [74] R. Feldman, B. Rosenfeld, M. Fresko, TEG—a hybrid approach to information extraction, *Knowledge and Information Systems* 9 (1) (2006) 1–18.
- [75] M. Vargas-Vera, E. Motta, J. Domingue, M. Lanzoni, A. Stutt, F. Ciravegna, MnM: ontology driven semi-automatic and automatic support for semantic mark-up. *Knowledge Engineering and Knowledge Management*, *Proceedings Lecture Notes in Artificial Intelligence* 2473 (2002) 379–391.
- [76] S. Liu, C.A. McMahon, M.J. Darlington, S.J. Culley, P.J. Wild, An automatic mark-up approach for structured document retrieval in engineering design, in: *Proceeding of the 4th International Conference on Manufacturing Research (ICMR'06)*, 2006, pp. 23–28.
- [77] T.S. Lee, D.W. Lee, J.H. Kim, Introduction of knowledge management system for technical support in construction industries, *Computational Intelligence and Security*, Part 2: *Proceedings of Lecture Notes in Artificial Intelligence* 3802 (2005) 749–756.
- [78] J.L. Hou, F.H. Lin, A hierarchical classification mechanism for organisation document management, *International Journal of Advanced Manufacturing Technology* 28 (3–4) (2006) 417–427.
- [79] C.P. Wei, C.S. Yang, H.W. Hsiao, T.H. Cheng, Combining preference- and content-based approaches for improving document-clustering effectiveness, *Information Processing and Management* 42 (2) (2006) 350–372.
- [80] C. McMahon, A. Lowe, S. Cully, M. Corderoy, R. Crossland, T. Shah, D. Stewart, Waypoint: an integrated search and retrieval system for engineering documents, *Journal of Computing and Information Science in Engineering* 4 (4) (2004) 329–338.
- [81] J. Furnkranz, Exploiting structural information for text classification on the WWW, in: *Proceedings of AIAD-99. The 3rd International Symposium on Advances in Intelligent Data Analysis*, 1999, pp. 487–498.
- [82] E. Glover, K. Tsioutsoulouklis, S. Lawrence, D. Pennock, G. Flake, Using web structure for classifying and describing web pages, in: *Proceedings of WWW 2002. The 11th International World Wide Web Conference*, 2002, pp. 562–569.
- [83] C.Y. Quek, Classification of World Wide Web documents. Senior Honour Thesis, School of Computer Science, Carnegie Mellon University, 1997.
- [84] Y. Yang, S. Slattery, R. Ghani, A study of approaches to hypertext categorisation, *Journal of Intelligent Information Systems* 18 (2–3) (2002) 219–241.
- [85] R. Ghani, S. Slattery, Y. Yang, Hypertext categorisation using hypertext patterns and meta-data, in: *Proceedings of ICML-01. The 18th International Conference on Machining Learning*, 2001, pp. 178–185.
- [86] J. Yi, N. Sundarsan, A classifier for semi-structured documents, in: *Proceedings of 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2000, pp. 340–344.
- [87] P. Denoyer Land Gallinari, Bayesian network model for semi-structured document classification, *Information Processing and Management* 40 (5) (2004) 807–827.
- [88] L. Denoyer, P. Gallinari, Using belief networks and Fisher kernels for structured document classification. *Knowledge Discovery in Database: PKDD*, *Proceedings of Lecture Notes in Artificial Intelligence* 2838 (2003) 120–131.
- [89] B. Piwowarski, G. Faure, P. Gallinari, Bayesian networks and INEX, in: *Proceedings of the first annual workshop of the INitiative for the Evaluation of XML retrieval (INEX)*, DELOS Workshop, Dagstuhl, Germany, 2002.
- [90] C.H. Caldas, L. Soibelman, J.W. Han, Automated classification of construction project documents, *Journal of Computing in Civil Engineering* 16 (4) (2002) 234–243.
- [91] G. Bordogna, G. Pasi, Controlling retrieval through a user adaptive representation of documents, *International Journal of Approximate Reasoning* 12 (1995) 317–339.
- [92] G. Bordogna, G. Pasi, Personalised indexing and retrieval of heterogeneous structured documents, *Information Retrieval* 8 (2) (2005) 301–318.
- [93] Y.J. Chen, Y.M. Chen, C.B. Wang, H.C. Chu, T.N. Tsai, Developing a multi-layer reference design retrieval technology for knowledge management in engineering design, *Expert Systems with Application* 29 (2005) 839–866.
- [94] M.C. Yang, W.H. Wood, M.R. Cutkosky, Design information retrieval: a thesauri-based approach for reuse of informal design information, *Engineering with Computers* 21 (2) (2005) 177–192.
- [95] S. Ahmed, Encouraging reuse of design knowledge: a method to index knowledge, *Design Studies* 26 (6) (2005) 565–592.
- [96] D. Egnor, R. Lord, Structured Information Retrieval using XML, 2000. <http://www.haifa.il.ibm/sigir00-xml/final-papers/Egnor/>.
- [97] S. Pradhan, K. Tanaka, Retrieving relevant portions from structured digital documents. *Database and Expert Systems, Applications Proceedings*. *Lecture Notes in Computer Science* 3180 (2004) 328–338.

- [98] S. Kim, B.T. Zhang, Web document retrieval by genetic learning of importance factors for HTML tags, in: Proceedings of the PRICAI-2000 Workshop on Text and Web Mining, 2000, pp. 13–23.
- [99] S. Kim, B.T. Zhang, Evolutionary learning of web-document structure for information retrieval, in: Proceedings of the 2001 Congress on Evolutionary Computation (CEC2001), 2001, pp. 1253–1260.
- [100] M. Agosti, F. Crestani, M. Melucci, On the use of information retrieval techniques for the automatic construction of hypertexts, *Information Processing and Management* 33 (2) (1997) 133–144.
- [101] Y. Chiaramella, Information retrieval and structured documents, *Lectures on Information Retrieval* (2001) 291–314.
- [102] F. Crestani, J. Vegas, P. Fuente, Design of a graphical user interface for focussed retrieval of structured documents, in: Proceedings of SPIRE'01. Symposium on String Proceeding and Information Retrieval, Laguna de San Rafael, Chile, (2001), pp. 246–249.
- [103] Y. Saidali, S. Adam, J.M. Ogier, E. Trupin, J. Labiche, Knowledge representation and acquisition for engineering document analysis graphics recognition: recent advances and perspectives, *Lecture Notes in Computer Science* 3088 (2003) 25–37.
- [104] I. Ruthven, M. Lalmas, A survey on the use of relevance feedback for information access systems, *Knowledge Engineering Review* 18 (2) (2003) 95–145.
- [105] I. Ruthven, M. Lalmas, K. Van Rijsbergen, Incorporating user search behaviour into relevance feedback, *Journal of the American Society for Information Science and Technology* 54 (6) (2003) 529–549.
- [106] M. Lalmas, T. Roelleke, Four-valued knowledge augmentation for representing structured documents, foundations of intelligent systems, *Proceedings Lecture Notes in Computer Science* 2366 (2002) 158–166.
- [107] M. Lalmas, A model for representing and retrieving heterogeneous structured documents based on evidential reasoning, *Computer Journal* 42 (7) (1999) 547–568.
- [108] K.C.K. Law, Y. Wang, H.H.S. Ip, A structured hypertext data model with versioning for engineering documents, *Multimedia Tools and Applications* 19 (3) (2003) 241–258.
- [109] M. Kim, P. Compton, Evolutionary document management and retrieval for specialised domains on the Web, *International Journal of Human-Computer Studies* 60 (2003) 201–241.
- [110] H.S. Shin, A storage and retrieval method of XML-based metadata in PVR environment, *IEEE Transactions on Consumer Electronics* 49 (4) (2003) 1136–1140.
- [111] P. Kotinurmi, H. Laesvuori, K. Jokinen, T. Soinien, Integrating design document management systems using the Rosettanet e\_business framework, in: Proceedings of the 6th International Conference on Enterprise Information Systems, vol. 4, Porto, Portugal, April 14–17, (2004), pp. 502–509.
- [112] K. Jokinen, J. Borgman, R. Sulonen, Common data model for design document exchanges in business-to-business networks. Proceedings of the 38th Hawaii International Conference on System Sciences, 2005. Available online at: <http://www.csd12.computer.org/comp/proceedings/hicss/2005/2268/04/22680095a.pdf>.
- [113] K. Thirunarayan, A. Berkovich, D.Z. Skol, An information extraction approach to reorganising and summarizing specifications, *Information and Software Technology* 47 (4) (2005) 215–232.
- [114] S.C. Brand, J. Morbach, M. Miatidis, M. Theiben, M. Jarke, W. Marquardt, Ontology-based information management in design process, in: The 9th International Symposium of Process Systems Engineering and 16th European Symposium on Computer Aided Process Engineering, Garmisch-Partenkirchen, Germany, (2006), pp. 9–13.
- [115] S.C. Wong, M.C. Crowder, G.B. Wills, On a service-oriented approach for an engineering knowledge desktop, WWW 2006, May 23–26, Edinburgh, Scotland, ACM 1-59593-323-9/06/0005.
- [116] S. Liu, C.A. McMahon, M.J. Darlington, S.J. Culley, P.J. Wild, A computational framework for retrieval of document fragments based on decomposition schemes in engineering information management, *Advanced Engineering Informatics* 20 (4) (2006) 401–413.
- [117] J. Reid, M. Lalmas, K. Finesilver, M. Hertzum, Best entry points for structured document retrieval—Part II. Types, usage and effectiveness, *Information Processing and Management* 42 (2006) 89–105.
- [118] P.J. Wild, C.A. McMahon, S.J. Culley, M.J. Darlington, S. Liu, Towards a method for profiling engineering documentation, in: Proceedings of the 9th International Conference of Design, Dubrovnik, May 15–18, 2006.
- [119] E.I. Neaga, J.A. Harding, A review of data mining techniques and software systems to improve business performance in extended manufacturing enterprises, *International Journal of Advanced Manufacturing Systems (IJAMS) Special Issue on Decision Engineering* 5 (2) (2002) 3–19.
- [120] J.A. Harding, M. Shahbaz, K. Srinivas, A. Kusiak, Data mining in manufacturing: a review, *Journal of Manufacturing Science and Engineering—Transactions of the ASME* 128 (4) (2006) 969–976.



**S. Liu**, Currently a Research Fellow who joined Computer Aided Design Centre in University of Strathclyde in 2006 after spending 3 years at the Innovative Manufacturing Research Centre in University of Bath. She obtained her PhD degree from Loughborough University, UK, in 2004, specialising in Information and Knowledge Management for Design and Manufacturing. Current research interests include study of information structures, Web technologies, information and knowledge systems, mark-up technology application to engineering domain, decision support systems, and integration of information systems.



**C.A. McMahon**, Professor of Engineering Design and Director of Bath Engineering Innovative Manufacturing Research Centre (IMRC). Principal Investigator of the EPSRC Grand Challenge project in *through-life information and knowledge management for product-service systems*. Chris's other research interests are predominantly concerned with the application of computers to all aspects of engineering design, including the organisation and delivery of design knowledge and information, multimedia and expert systems, etc.



**S.J. Culley**, Professor of Design Technology, Theme leader of Design Information And Knowledge (DIAK) in Bath Engineering Innovative Manufacturing Research Centre (IMRC). Steve has wide expertise in engineering design including Components Selection, Feedback and Decision Making, Electronic Catalogue, Optimisation, Parametric representation, Constraint and Data propagation Design Information, Designer requirements and access methods.