ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

State of The Art Content Mining Using SCAN Technology

Dr. Anuj Kumar Parashar [1], Er. Dev Kant Tyagi[2]

Assistant Professor, Department of Computer Science & Engineering, FET Agra College, Agra

Uttar Pradesh, India[1]

Research Scholar, Department of Computer Science & Engineering, FET Agra College, Agra

Uttar Pradesh, India[2]

ABSTRACT: Smart Content is same like Semantic Web Service Composition (SWSC) which is defined as generating the aggregated service by integration of independent available component services for satisfying the client each request that cannot be satisfied by any available single service. Thus, we can say that in most cases semantic web service also not able to satisfy client request by single service component. That's why we need the service composition for generating aggregation of service components according to the requested task.

The basic and main objective of project is to provide more efficient approaches for composition on semantic mobile services. The thesis also aims at providing a theoretically complete classification system for various composition approaches. Thesis presents framework for tag based Smart Content service composition and its various models also. Here we try to search heterogeneous contents present in mobile's SD Card but not discoverable due to the variety of contents and various file types. We try to develop a tag based repository that can be searched for desired keyword which finds file(s) having that keyword. The difficulty to achieve the result is due to large number of file types as for each extension we have to code differently. This thesis presents an implementation for Smart Content Tag Repository and mainly focuses Android Mobiles. We have also presented a comparative study of various smart content aggregation and navigation technology on various platforms and finally thesis presents an implementation of this technology on a widely used Android Operating System for Smart Phones.

KEYWORDS: Tag, Smart Content, Aggregation, SCAN, Paper Submission.

I. INTRODUCTION

The Smart Content

Growing amount of information dispersed across different sources is an increasing problem of state-of-the-art information management. To be solved effectively, it demands new approaches and tools, strongly focused on the content semantics and supported by automation and intelligence. (1)

Smart Content can be seen as the "product bridge" between "meaningful use" data analytics business and "actionable expert opinion" business -- the data supplying the "what" analytics and semantics to:

a) Creation of new products with contemporary user experience also for services.
b) Deliver an answer not a result (drive better usability of existing products, increase discoverability and content access, provides unified access or interoperability, integration with search).
c) For realizing latest dynamic monetization models.

SCAN (Smart Content Aggregation and Navigation) technology combines semantic integration, search and natural language processing for intelligence of document management in the age of information overload. In order to provide new, improved experience for knowledge workers, the technology addresses a broad range of major issues and challenges of state-of-the-art information management:

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

- The information is dispersed across different heterogeneous sources and locked into numerous application-specific formats. It sets a barrier for a high-level overview of existing knowledge base and find-ability of semantically related information pieces.
- The structure of information is mostly dictated by a physical storage technology (an example – files and folders in the file system), while more high-level, semantic structures are needed.
- There is no uniform way to describe and annotate the content resources, supported consistently across a whole system. Even if a document format supports metadata, it does not contribute into overall information management environment and is mostly useless outside an application that uses this format.
- Despite of the computers are smartest of all artificial things now, there is no pro-active help in the content classification and information modeling from their side. The computers role in information management still to be the passive storage and processing systems.

you still can store your documents where you are used to, send and receive emails, bookmark your favorite websites and so on. SCAN does not touch your content either, but adds a layer upon it to turn your content to the *smart content*– that is, the content that *knows something about itself*.

An integrated approach

SCAN approach is based on an idea that information overload is a complex of problems and no single magic bullet exists to solve them effectively. Thus, there is a synthesis of few different techniques aimed at specific aspects of the problem.

1.1 Content aggregation

SCAN erases the boundaries put on information by different storage systems. It links the information items from multiple sources and of different formats into a seamless digital library, where they can be categorized, annotated navigated and searched by a uniform way. This provides a homogeneous searchable and explorable semantic info space where files, emails ,web-pages, other content items are equal documents organized by their natural semantic properties.

The component architecture makes the technology agnostic of specific types of sources (local or network file systems, web-sites etc.) and of the document formats (MS Office, PDF, HTML …). A number of those types of sources and formats can be supported via integration of the components for a specific business application.

1.2 Tagging

Tagging is the easiest and intuitive way of information modeling and organization of the documents collection similarly to the popular services like delicious or Flicker. Tags makes the manual document tagging as simple as selecting the tags from the suggested candidates and Also, a user can make entrust the process of tagging whole to the system, so that the documents would be tagged automatically with the relevant terms.

1.3 Text analysis and concept extraction

SCAN brings the power of automated text mining (2) and natural language processing to discover document semantics by extracting the valuable terms and their patterns from the document content. It makes possible to identify what the document is about and how it relates to others.

Text analysis greatly simplifies the process of tagging. It helps a user to pick the most relevant terms identifying a document and assign them as the document tags. Text analysis is also underlying the advanced semantic search functionality like finding the documents by similarity (pattern search) and associative guided search based on system suggestions.

1.4 Metadata and facet navigation

SCAN provides a rich set of metadata properties associated with the documents, including document title, description or annotation, writer, creation date and others. The properties are set automatically on a document adding and can be quickly edited later.

Metadata properties can be used in the structured search to find the documents matching specified criteria.

The documents content is indexed for search – either unstructured full-text search or more complex, structured queries both on text and metadata properties.

Text analysis functionality is driven Advanced semantic search techniques are driven. After any search request is performed, results are analyzed to build a “see also” terms list for associative search. It provides step-by-step

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

exploration of an area of interest following the system recommendations. Pattern search explores semantic compatibilities between different documents and allows finding the documents conceptually similar to a subject of a given one. (3)

1.5 Smart Content Readiness Service

It will engage with organizations on a consulting basis to identify:

- The business drivers where smart content will ensure competitive advantage when distributing business information to customers and stakeholders
- The technologies, tools, and skills required to componentized content, and target distribution to various audiences using multiple devices
- The operational roles and governance needed to support smart content development and deployment across an organization
- The implementation planning strategies and challenges to upgrade content and creation and delivery environments

1.6 Smart Content Technology landscape

What it is and the benefits it can bring to an organization include flexible, dynamic assembly for delivery to different audiences, search optimization to improve customer experience, and improvements for distributed collaboration. It might help to better understand the technology landscape involved in creating and delivering smart content."(6)

The figure below illustrates the technology landscape for smart content. At the center are fundamental XML technologies for creating modular content, managing it as discrete chunks (with or without a formal content management system), sometimes referred to as Singe Source Publishing (SSP) systems.

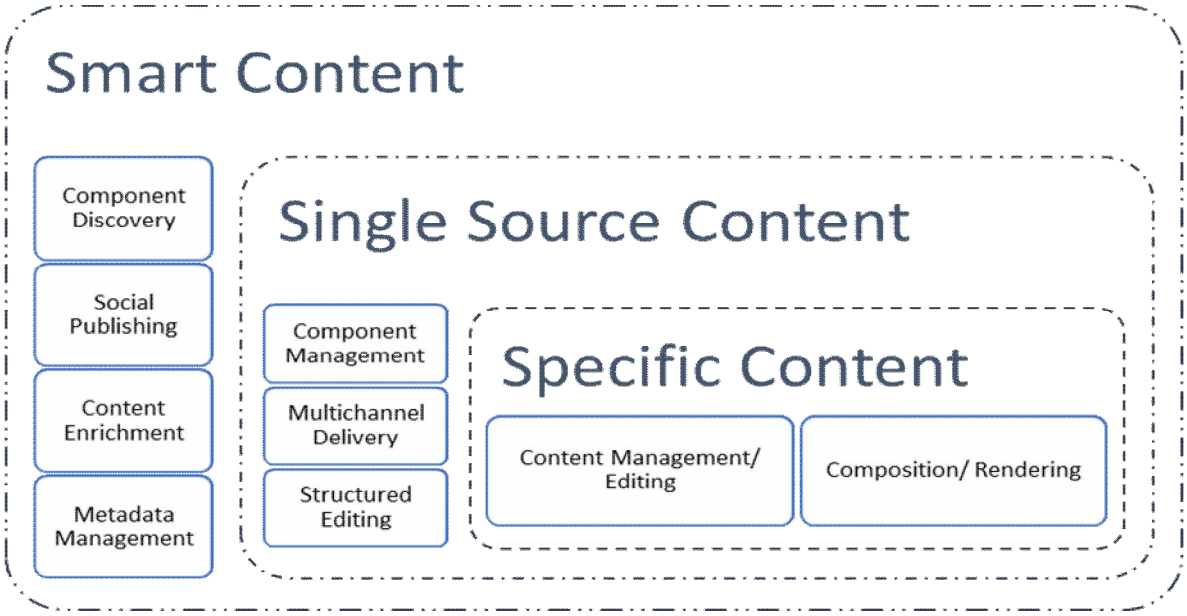

Figure 1.1Smart Content Layer Above the Single Source Publishing(SSP) Systems

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

II. PROPOSED SYSTEM ARCHITECTURE

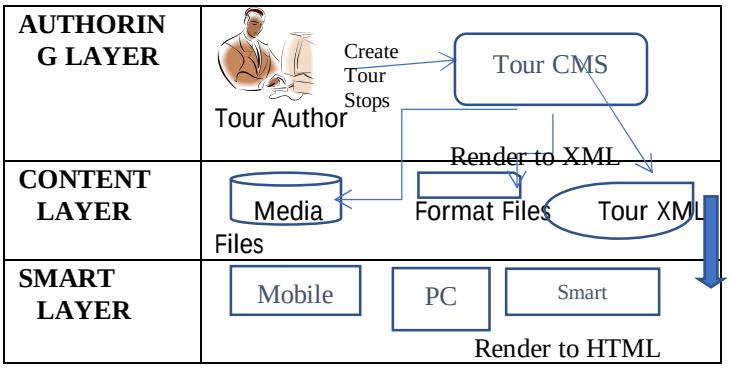

Figure 2.1 Content Management System (CMS)

Here's the basic idea. Content authors don't want to worry about the technical details about how to structure and validate XML or HTML documents, they just want to be able to focus on the task of creating great content. Authors also need to be able to preview the tour and tour stops to verify that the correct media is being used and that navigational connection are correct, etc.

This could be done by deploying to the final device and checking there, but in reality, that process is a bit time-consuming and will slow the content development process down considerably.

A CMS seems to be the perfect tool for this job. The creation of a back-end Content Management System is proposed with a restricted set of content types reflecting the different types of tour content that can be represented by the TourML XML schema. Such a CMS can be used by authors to create descriptive content, upload media, and organize the flow and navigation of a tour. CMS functionality can do nice things like handle all the resizing/cropping of images, transcode video and audio to a standard set of formats, present structured input fields for the creation of repeatable content. A CMS can also provide a simple web-based rendering of the tour content and navigation allowing the content author to "try out" and tweak the tour prior to bundling a platform independent content package of the tour.(11)

Once the author is satisfied with the tour, they will choose to "Publish" or "Export". Publishing a tour will issue a number of callbacks which will be executed by the CMS. Content types from the CMS will be traversed to create an XML instance of this tour using the TourML schema. In addition, any media assets associated with the stops on a tour (i.e. audio files, video files, and uploaded images) will be bundled together into a zip file or other type of package that can be deployed with the TourML document to the application device platform. At the conclusion of the Publishing process, the TourML document and the Media bundle should represent a complete and portable version of the tour consisting of all the information needed to reconstruct this tour on any number of devices.

Deploying a tour to a mobile device may take a number of paths depending on the desired platform and use case. Museums may choose to copy the tour bundle to devices for "offline" use where a wireless network is not available. Museums might choose to host this tour content on a web server inside or outside of their museum, and then render this content into style sheets branded for their museum, and/or as mobile web pages designed for small-screen rendering.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

III. METHODOLOGY

3.1 Algorithm Development or Formulation

An effective searching algorithm is used to tag files using the occurrences of the words in the specific file. Top 10 data mining algorithms identified by the IEEE International Conference on Data Mining (ICDM) in December 2006: C4.5, k-Means, SVM, Apriority, EM, PageRank, adaBoost, kNN, Naive Bayes, and CART. Derived algorithm as follows:(18)

ALGORITHM

1. Put the words you want to search for in a hash table, with the words as keys and the values initialized to 0.
2. Iterate over the words of the text, each time checking to see if the word is a key in the hash table, if so, then increment the value for that key.
3. Iterate over the hash table finding the values which are non-zero, the keys for these are matched words, the values are the counts.

Runs in **O (N+M)** where N is the number of words you're searching *for* and M is the number of words you're searching *through*.

3.2 Algorithm Implementation

1. Open file in read mode.
2. Start reading the words from Start-Of-File (SOF) excluding the common words (like is, are, am, the, then, has, have, had, will, shall, and, or, etc.) and conjunctions.
3. Store the keywords in Dictionary while End-Of-File (EOF) is reached with the following conditions:
    a. If key already exists increment the value by 1
    b. If key does not exist initialize value to 0
4. List the keys having top 10 values.
5. Store these keys in project repository as tags.
6. Save the file location corresponding to keys.

3.2.1 Model Used Work Flow

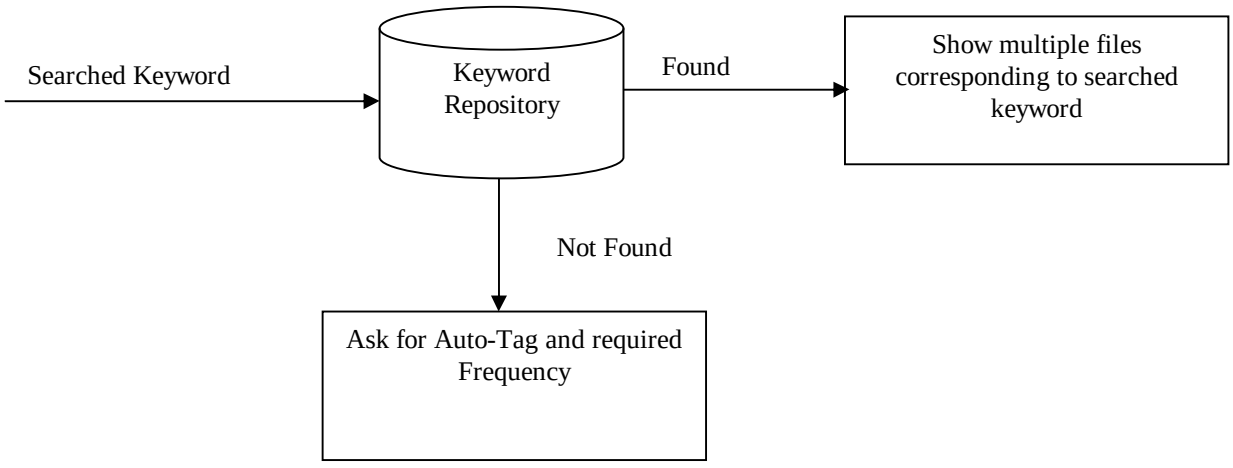

Figure 3.2 Tag Search Process

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

Tagging multiple files manually requires repetition of process.

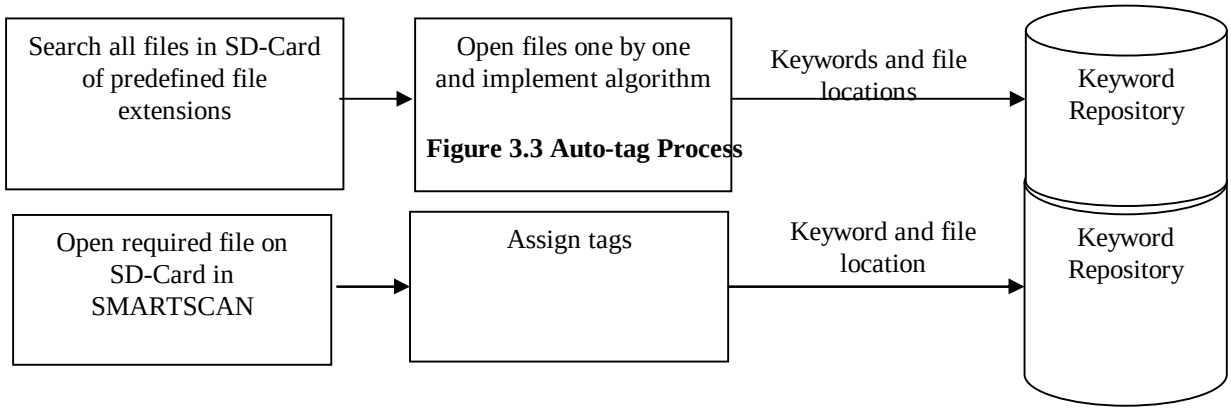

Figure 3.3 Auto-tag Process

Figure 3.4 Manual-tag Process

IV. CONCLUSIONS

It can be concluded from various literature survey and case studies that the work can be integrated with semantic web, ontology, cloud, and other business structures to manage and fetch the documents on the go and there is lesser need of documentation at the end. Some future scope has been discussed below:

4.1 PLUG-INS FOR NEW DOCUMENT FORMATS

SCAN platform can be easily extended with plug-ins for new document formats, document locations (RSS feeds, web-sites, e-mail, etc.) and language analyzers. Whole new areas of functionality can be added with user interface extensions. An example of such extensions is the plug-in to browse the repository with a calendar (grouping the documents by their creation dates).
Another text analysis application may be searching the documents similar to a specific one (search by pattern).
What's on blog now a days gives the idea of the research works going on this domain and integrating other domains and Content Technologies track? Some of those topics cover: standards, integration, content migration, search, open source, and relevant consumer technologies. The future work may contain:
1. Multi-lingual technologies and applications, XML, standards, integration, content migration, hand held devices, searching, open source, SaaS, semantic technologies, social software, SharePoint, XBRL, and relevant consumer technologies.
2. Business-oriented applications of the SCAN technology; design and development of the server version for small and medium enterprise networks.
3. Integration with Cloud Computing; by implementing and integrating their ideas and works with content technology.

4.2 EXTENSIBILITY TO CLOUD APPLICATION

Digital marketers, senior IT folks and business analysts faced with the decision to deploy these technologies outside the server room. In it set out to answer the following questions:(9)

- **What do we mean by the cloud?** There is a great deal of hype, sales, and marketing messaging around "the cloud." We explore what it really is and the opportunities it represents for digital marketers.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

- **What are the deployment options when working with a cloud platform partner?** The decision around deploying to the cloud is not always a binary choice to host in the server room or not. We look at possible solution architecture options and the benefits of each.
- **What do organizations need to look for in a WEM solution in the cloud?** If deploying into the cloud is an attractive option for an organization, we consider the key attributes that organizations should build into their selection criteria when choosing a solution.

Outsell's Content Technology Consulting services leverage Outsell's information technology consulting for:(29)

- Content Strategy Development
- Technology Assessments
- Content Technology Vendor Selection and Evaluation
- Product IT and Technology Function Benchmarks and Best Practices

Cloud Content Management adds a low-cost, complementary layer to existing ECM systems that enables workers to better collaborate securely around active content. In the future, Cloud Content Management functionality could potentially expand to also address the capture and archival stages of the content lifecycle.
On the front end of the content lifecycle, the ability to create digital documents using software hosted in the cloud has existed for several years now (e.g. Google and ZohoDocs). More recently, technologies have been introduced that allow multiple authors to collaborate simultaneously on the same document (i.e. Google Wave.) In the future, Cloud Content Management platforms could provide services that would allow users to scan documents, shoot photos, and record audio or video then tag and send the resulting digital files to a managed, active repository hosted in the cloud. In fact, these technologies already exist as mobile applications on smartphones; the next logical step would be to provide that functionality on other computing devices. (10)

At the back end of the content lifecycle, Cloud Content Management's existing search tools could be used for reliable eDiscovery of content in the active repository that has been locked to further editing or sharing by the file's owner. Perhaps that locking activity could be automated by adding system rules based on established corporate records declaration policies. Locked files could also be deleted automatically from the Cloud Content Management system based on records retention requirements. Those rules could be applied manually when the content was created or uploaded into the repository. A comprehensive approach such as this would eliminate the need to have an archival area into which files designated as records are moved and, eventually, removed and destroyed.

Many of these potential expansions of Cloud Content Management functionality to address content capture and retention/disposal requirements would be especially welcomed by small and medium organizations, who need to address those pieces of the content lifecycle, but in a lightweight manner. Those organizations do not have the resources to invest in the acquisition, deployment, and operation of an Enterprise Content Management system, but could be very well served by a less expensive, more agile Cloud Content Management platform.

There is, of course, the potential for Cloud Content Management solutions to provide even more process-based control of active content. Common content collaboration patterns now treated as ad hoc workflows could be codified into standardized, yet flexible rule-based processes. For example, if an individual could subscribe to a specific content tag, she could be automatically notified and served a link whenever a new, publicly-accessible file bearing that tag appeared in the Cloud Content Management system. In effect, this would automate the sharing of that file, replacing the all-too-common practice of manually emailing the file to a poorly constructed and maintained distribution list. It is important, however, that future work to automate content collaboration patterns not be done at the expense of the simplicity and ease of use that is one of the primary benefits of Cloud Content Management today.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

4.3 SMART CONTENT IN PUBLISHING

Smart content remains a work in progress. Geoffrey Bock expects to develop the prescriptive road map in the months ahead. Here's a quick take on where he is right now.(30)

- For publishers, it's all about transforming the publishing paradigm through content enrichment – defining the appropriate level of granularity and then adding the semantic metadata for automated processing.
- For application developers, it's all about getting the information architecture right and ensuring that it's extensible. There needs to be sensible storage, the right editing and management tools, multiple methods for organizing content, as well as a flexible rendering and production environment.

For business leaders and decision makers, there needs to be an upfront investment in the right set of content technologies that will increase profits, reduce operating costs, and mitigate risks.

REFERENCES

[1]. *Web Content Management - WCM.* Clea. Boston : Gilbane Conference, 2014.
[2]. *Text Mining with Information Extraction.* Mooney, Raymond J. and Nahm, Un Yong. Austin : University of Texas, 2013.
[3]. Waldt, Dale.*Next Generation ECM for Mission Critical Applications: Open Source ECM Capabilities and Opportunities.* Boston : Gilbane Community, 2010.
[4]. Seth Grimes.*Six Definitions of Smart Contents.* 2010.
[5]. Geoffrey G, Bock, Waldt, Dale and Mary, Laplante.*Smart Content in the Enterprise: Next Generation XML Applications.*
[6]. *Understanding the Smart Content Technology Landscape.* Waldt, Dale. s.l. : CMSWire, 2010. CMS Advisor.
[7]. *The Gilbane Conference.* Waldt, Dale. 2010.
[8]. Bock, Geoffrey and Waldt, Dale.*Managing Content for Continuous Learning at Autodesk.* s.l. : The Gilbane Group, 2011.
[9]. *Taking Online Engagement to the Cloud.* Laplante, Mary. s.l. : The Gilbane Group, 2011.
[10]. *Cloud Content Management: Facilitating Controlled Sharing of Active Content.* Hawes, Larry. s.l. : The Gilbane Group, 2010.
[11]. White, Martin.*The content management handbook.* s.l. : Facet Publishing, 2005.
[12]. Yizhou Sun and Jiawei Han.*Mining Heterogeneous Information Networks: A Structural Analysis Approach.*
[13]. George, David.*Understanding Structural and Semantic Heterogeneity in the Context of Database Schema Integration.* 2011.
[14]. *Conception of Information Systems.* Aberer, Karl. s.l. : Laboratoire de systèmes d'informations rèpartis, 2003, Laboratoire de systèmes d'informations rèpartis.
[15]. *Platforms: Experimenting a Service Based Connectivity between Adaptable Android, WComp and OpenORB.* Monfort, Valerie and Cherif, Sihem. 2012.
[16]. Trippe, Bill.*Component Content Management.* s.l. : Gilbane Group, 2008.
[17]. *Information Mining.* Rudolf Kruse. s.l. : EUSFLAT Conference, 2001.
[18]. XindongWu, et al., et al.*Top 10 algorithms in data mining.* Verlag-London : Springer, 2007.
[19]. Marc Strohlein. Content Immediacy: The New Marketing Imperative. *Guidance on content Strategies, Practices and Technologies.* February 2012.
[20]. *Are You Leveraging All the Mobile Technologies Required for Competitive Mobile Engagement?* Marc Strohlin, Frank Schneider and Luke Barton. December 3, 2013.
[21]. Werner Behrendt, et al., et al.*EP2010: Dossier on Smart Content* . September 2003.
[22]. Rockley, Ann, Kostur, Pamela and Manning, Steve.*Managing Enterprise Content: A Unified Content Strategy.* s.l. : New Riders, 2003.
[23]. *State of the ECM Industry.* s.l. : AIIM Industry Watch 2009.
[24]. *Content Management Interoperability Services (CMIS).* Waldt, Dale. s.l. : The Gilbane Group, 2009.
[25]. *Towards Smart Publishing at IBM.* Bock, Geoffrey and Waldt, Dale. s.l. : The Gilbane Group, 2010.
[26]. Paxhia, Steve and Rosenblatt, Bil.*Digital Magazine and Newspaper Edition.* s.l. : Gilbane Group, 2008.
[27]. *Documenting Semiconductor Devices at IBM.* Bock, Geoffrey. s.l. : The Gilbane Group, 2010.
[28]. *Linked Data in Pearson- The Asset Enrichment Process.* Solomon, Madi Weland and Johnson, Marlowe. London : s.n., 2013.
[29]. Mary Laplante. Smart Approaches To Managing Mobile Learning Content. s.l. : Outsell Inc, 2011.
[30]. *What's Next with Smart Content.* Bock, Geoffrey. Boston : s.n., 2010. The Gilbane Confere.
[31]. *Review on Web Content Mining Techniques.* IJCA(0975-8887). Amity University : Volume 118-No. 18.,May 2015.
[32]. *Literature survey in Web Content Mining.* IJRITCC. Trichy, Tamilnadu, India : Volume 4,Issue: 10.,Oct 2016.
[33]. *.A Review of Trends in easearch on Web Content Mining* IJMTER. Trichy, Tamilnadu, India : Volume 3,Issue: 10.,Oct 2016, ISSN(Online):2349-9745

BIOGRAPHY

Dr. Anuj Kumar Parashar is a Assistant Professor in the Computer Science Department, Faculty of Engineering & Technology, A.K.T.U, Lucknow. He received Ph.D. degree in 2014, UP, India. His research interests are Differential Evolution, Multi-Objective Optimization, Content Mining Technology.

Er. Dev Kant Tyagi is a Research Scholar in the Computer Science Department, Faculty of Engineering & Technology, A.K.T.U, Lucknow. He is about to receive Master of Technology degree in 2017 from A.K.T.U., Lucknow, UP, India. His research interests are Android and java technology.