STATE OF THE ART CONTENT MINING USING SCAN TECHNOLOGY

Nishant Mudgal[1], Mr. Sambhav Agarwal[2]

[1]M.Tech, Computer Science and Engineering, SR Institute of Management & Technology, Lucknow, India
[2]Associate Professor, Computer Science and Engineering, SR Institute of Management & Technology, Lucknow

---

Abstract - *This research paper explores the state of the art in content mining using SCAN (specific context-aware network) technology. Content mining, the process of extracting valuable insights and knowledge from vast amounts of textual and multimedia data, has gained significant attention in various domains. However, existing content mining techniques often face challenges in accurately capturing contextual information and extracting relevant content. The SCAN technology aims to address these limitations by incorporating context-awareness into the mining process.*

This paper begins by providing an overview of the existing content mining approaches and their shortcomings. It then introduces the SCAN technology, highlighting its key features and advantages. SCAN leverages advanced natural language processing (NLP) techniques, machine learning algorithms, and semantic analysis to enhance context understanding and extract meaningful information from diverse content sources.

The research paper discusses various applications where SCAN technology can be effectively utilized, such as sentiment analysis, recommendation systems, information retrieval, and knowledge discovery. It explores the benefits of SCAN in improving the accuracy and relevance of mining results, leading to enhanced decision-making and user experiences.

Key Words: content mining, SCAN technology, context-awareness, natural language processing (NLP), machine learning, semantic analysis, sentiment analysis.

1. INTRODUCTION

The concept of smart content has evolved alongside advancements in technology and the increasing demand for personalized digital experiences. Here is a brief overview of the history of smart content:

1. Early Personalization Efforts: In the early days of the internet, personalization efforts were limited to basic customization options such as choosing preferences for website colors or layouts. These simple personalization features aimed to provide a more tailored experience but lacked the sophistication and intelligence seen in modern smart content.

2. Rise of Recommendation Systems: With the growth of e-commerce and online services, recommendation systems emerged as a key component of smart content. Websites started using collaborative filtering and data analysis techniques to suggest products, movies, music, or articles based on user behavior and preferences. Amazon's "Customers who bought this also bought" feature is a classic example of early recommendation systems.

3. Dynamic Content Generation: As web technologies advanced, content management systems (CMS) became more capable of generating dynamic content. This allowed websites to display different content to users based on various factors like location, demographics, and browsing history. Websites began to use cookies and user profiles to deliver personalized content, advertisements, or promotions.

4. Big Data and Machine Learning: The advent of big data and machine learning techniques further revolutionized smart content. With the ability to process and analyze vast amounts of user data, algorithms could learn and predict user preferences with greater accuracy. This led to more personalized experiences across various platforms and applications.

5. Internet of Things (IoT) Integration: The integration of smart devices and IoT technologies expanded the scope of smart content beyond websites and applications. Connected devices like smart speakers, wearables, and home automation systems began offering personalized content and recommendations based on user behavior, location, and sensor data.

6. Artificial Intelligence (AI) and Natural Language Processing (NLP): AI and NLP technologies have played a significant role in advancing smart content capabilities. Chatbots and virtual assistants, powered by AI, can understand and respond to user queries in a conversational manner. Natural language understanding and sentiment analysis enable better understanding of user intent and provide more relevant content recommendations.

7. Context-Aware and Real-Time Personalization: The latest developments in smart content focus on real-time, context-aware personalization. By considering real-time data such as user location, time of day,

device type, and environmental factors, content can be dynamically adapted to meet immediate user needs. This includes displaying relevant offers, providing location-based information, or adjusting content presentation based on screen size or orientation.

Smart content refers to digital content that is dynamic, adaptive, and personalized to meet the needs and preferences of individual users. It goes beyond static information by leveraging technology to provide a more interactive and tailored user experience.

In simple terms, smart content is like a chameleon that can change its appearance and behavior based on who is interacting with it. It takes into account factors such as user demographics, interests, browsing history, and location to deliver content that is relevant and engaging.

For example, imagine visiting a website that displays different content depending on whether you're a new visitor or a returning customer. The smart content system may showcase personalized recommendations, targeted advertisements, or customized product suggestions based on your previous interactions with the website.

Smart content can also adapt to different devices and screen sizes. It ensures that the content is appropriately displayed and optimized for various platforms such as desktop computers, smartphones, and tablets. This flexibility allows users to have a consistent and seamless experience across different devices.

1.1. AN INTEGRATED APPROACH AND CONTENT AGGREGATION

An integrated approach to content aggregation refers to the combination of various techniques and strategies to gather and organize content from multiple sources into a unified and cohesive experience. It involves the collection, filtering, and presentation of relevant information from diverse channels and platforms.

The process of content aggregation typically involves the following steps:

1. Source Identification: Identifying and selecting relevant sources of content is the first step. These sources can include websites, blogs, social media platforms, news outlets, and other content repositories.

2. Content Collection: Once the sources are identified, the content aggregation system collects data from these sources. This can be done through automated web scraping, RSS feeds, APIs (Application Programming Interfaces), or manual curation.

3. Content Filtering: Content filtering is crucial to ensure that only relevant and high-quality content is included in the aggregation. Filtering techniques can involve the use of keywords, metadata analysis, sentiment analysis, or machine learning algorithms to identify and prioritize content based on specific criteria.

4. Content Integration: The collected and filtered content is then integrated into a unified system or platform. This integration can involve combining text, images, videos, and other media formats into a cohesive display.

5. Content Presentation: The aggregated content is presented to users in a user-friendly and easily consumable format. This can be achieved through the use of personalized dashboards, news feeds, topic-based categorization, or customized layouts.

An integrated approach to content aggregation ensures that users have access to a wide range of relevant and up-to-date information from various sources without the need to visit multiple platforms individually. It simplifies the process of content discovery, saves time, and provides users with a comprehensive view of their interests or chosen topics.

Furthermore, integrating different types of content, such as news articles, blog posts, social media updates, and multimedia, enhances the richness and diversity of the aggregated content. It allows users to engage with a variety of content formats and media sources within a single platform.

2. TEXT ANALYSIS

Text analysis and concept extraction are techniques used to analyze textual data and extract meaningful information and concepts from it. These techniques are widely employed in various fields, including natural language processing (NLP), information retrieval, sentiment analysis, and knowledge discovery. Here's a brief explanation of text analysis and concept extraction:

1. Text Preprocessing: This involves cleaning and preparing the text data by removing noise, punctuation, stopwords, and performing tasks such as tokenization (breaking text into individual words or sentences), stemming (reducing words to their base form), and lemmatization (converting words to their base dictionary form).

2. Sentiment Analysis: Sentiment analysis determines the emotional tone expressed in a piece of text, classifying it as positive, negative, or neutral. It helps understand the sentiment of customers, users, or the general public towards a product, service, or topic.

3. Named Entity Recognition (NER): NER identifies and classifies named entities in text, such as person names, locations, organizations, dates, and other relevant entities. It helps extract specific information and can be useful for information retrieval or knowledge extraction tasks.

4. Topic Modeling: Topic modeling techniques, such as Latent Dirichlet Allocation (LDA), identify and extract underlying themes or topics from a collection of documents. It allows for a high-level understanding of the main concepts present in the text corpus.

2.1. CONCEPT EXTRACTION

Concept extraction focuses on identifying and extracting specific concepts or entities from text. It goes beyond individual words and aims to capture meaningful units of information. Concept extraction techniques include:

1. Named Entity Recognition: As mentioned earlier, NER identifies and extracts named entities from text, which can be considered as concepts.

2. Relation Extraction: Relation extraction aims to identify and extract relationships between entities mentioned in the text. For example, extracting relationships between people and organizations or identifying the subject and object of a sentence.

3. Keyphrase Extraction: Keyphrase extraction techniques identify and extract important phrases or terms from text that represent the main concepts or topics discussed. These keyphrases can be useful for summarization, indexing, or information retrieval purposes.

4. Ontology or Knowledge Graph-based Extraction: This approach involves mapping entities and their relationships to a predefined ontology or knowledge graph, enabling the extraction of structured and semantically rich concepts.

3. METADATA AND FACET NAVIGATION IN SCAN

Metadata refers to descriptive information or data that provides details about the content itself. It includes attributes such as title, author, publication date, keywords, tags, and other relevant information associated with a particular piece of content. In SCAN, metadata plays a crucial role in understanding and contextualizing the content.

By analyzing metadata, SCAN can gain insights into the characteristics and properties of the content. It helps in identifying relevant content based on specific criteria or user preferences. For example, metadata can be used to filter and prioritize content based on factors like publication date, source credibility, or relevance to a particular topic.

Additionally, metadata can provide valuable context for content aggregation and recommendation systems. By leveraging metadata, SCAN can tailor content suggestions based on user interests, historical interactions, or other relevant metadata attributes. Metadata enhances the understanding and relevance of content in the SCAN ecosystem.

Facet navigation, also known as faceted search or faceted browsing, is a user interface technique that allows users to explore and filter content based on multiple dimensions or facets. In SCAN, facet navigation provides users with flexible and dynamic options to navigate and refine their content exploration.

Facets represent different dimensions or attributes associated with the content. These facets can include metadata attributes such as author, publication date, category, source, sentiment, or any other relevant criteria. By presenting these facets to users, SCAN enables them to navigate and filter content based on their preferences.

Facet navigation allows users to dynamically select or combine facets to refine their content exploration. For example, a user can choose to explore content from a specific author within a particular time range or select multiple categories of interest. This interactive and iterative process empowers users to customize their content discovery experience within the SCAN environment.

Facet navigation in SCAN enhances the user's control and flexibility in finding content that matches their specific requirements or interests. It helps in narrowing down the content space, reducing information overload, and improving the overall user experience by providing relevant and context-aware content recommendations.

3.1. METADATA MANAGEMENT

Metadata management refers to the processes, practices, and technologies involved in the collection, organization, storage, and maintenance of metadata. Metadata, in this context, refers to structured information that provides descriptive and contextual details about data or content.

Effective metadata management is crucial for ensuring the accuracy, consistency, and usability of data and content within an organization or system. Here are some key aspects of metadata management:

1. Metadata Collection: Metadata collection involves capturing and recording relevant metadata about data or content at various stages of its lifecycle. This can include information such as data source, creation date, author, data type, format, and other attributes that describe the content or data's characteristics and context.

2. Metadata Storage and Organization: Metadata needs to be stored in a structured manner to enable efficient retrieval and management. This typically involves creating a metadata repository or database where metadata is stored, organized, and indexed. Different metadata standards and schemas, such as Dublin Core, MARC, or industry-specific schemas, may be used to ensure consistency and interoperability.

3. Metadata Integration: In many organizations, data and content are stored in multiple systems and formats. Metadata integration involves harmonizing and mapping metadata from different sources and systems, enabling seamless access and interoperability across these systems. Integration can be achieved through data integration tools, APIs, or data modeling techniques.

4. Metadata Quality Assurance: Metadata quality is essential for accurate and meaningful data interpretation. Metadata management includes processes and controls to ensure metadata accuracy, completeness, consistency, and relevance. This may involve data profiling, validation, and data governance practices to maintain high-quality metadata.

5. Metadata Search and Retrieval: An important aspect of metadata management is enabling efficient search and retrieval of data or content based on specific criteria. This involves developing search interfaces and tools that leverage metadata attributes to facilitate quick and accurate retrieval of relevant information.

6. Metadata Governance and Security: Metadata governance involves establishing policies, standards, and procedures for metadata management. It ensures that metadata is governed in a consistent and controlled manner, promoting data integrity, security, and compliance with regulatory requirements. Metadata security measures may include access controls, encryption, and data privacy practices.

7. Metadata Lifecycle Management: Metadata, like data and content, goes through a lifecycle. Metadata management encompasses activities throughout this lifecycle, including metadata creation, modification, archiving, and retirement. It ensures that metadata remains relevant, up-to-date, and aligned with the evolving needs of the organization.

Effective metadata management provides numerous benefits, including improved data and content discoverability, enhanced data integration, better decision-making, and increased data quality and consistency. It forms the foundation for efficient data management, data governance, and information management practices within organizations.

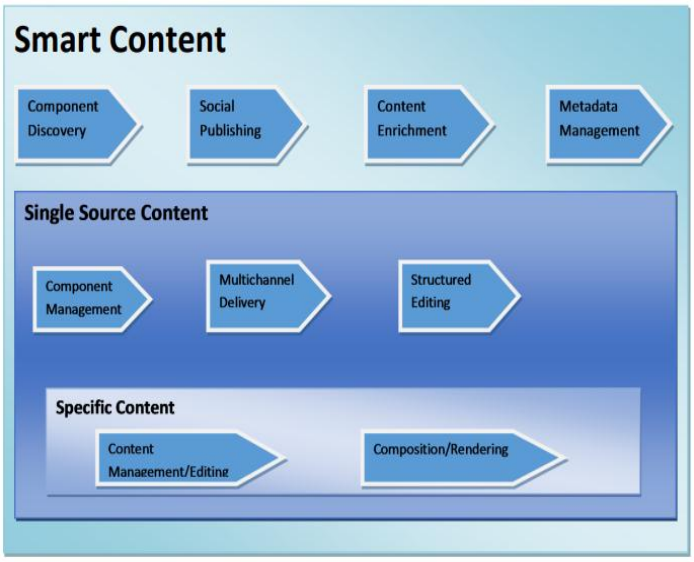

Figure-1: Smart Content.

4. PROPOSED SYSTEM ARCHITECTURE

The gist of it is as follows. Authors of content would rather concentrate on their work than on the technicalities of document structure and validation in markup languages like XML or HTML. The tour and its stops must be previewed for the author to ensure that the appropriate media is being utilised, that navigational connections are functioning properly, etc. Deploying to the target device and doing quality assurance checks there is one option, but doing so is time-consuming and would slow down content creation. Museums can take advantage of the iPhone's browser-based user interface by either developing an iPhone App that can interpret the TourML schema and reference media files stored locally on the device, or by using webkitand dashcode based tools like iWebKit and JQTouch. The application layer is in charge of the finer elements of the user interface, such as the appearance and feel, the graphical design of the tour, and the need for wireless connection. There will be a wide variety of implementations of this application layer among museums, however Museums-To-Go might offer some reference implementations that can read and interpret the TourML standard. The iPhone app, a mobile web-based app, an iPhone app that uses a dash code, etc., are all viable options for these examples.

5. METHODOLOGY

The information you need may be stored in a variety of formats, making data retrieval and mining more difficult. The following are only a few of the numerous serious concerns:

1. Determining which software development kit (SDK) will aid me in delivering a suitable platform. Determining the best programming language for maximising my impact on smart phones.

2. Deciding on the most appropriate algorithm.

3. Deciding what kind of records will be considered.

5.1. METHODOLOGY USED

Manual Tagging: When analysing a downloaded file, if the user decides that it would be beneficial, they may add tags to it with relevant keywords. These tags can then be used to find the item in the future. Document Exploration Using Metadata: In the future, if you wish to search for a file using a certain term, you may simply provide that phrase to the programme, and it will either manually search for the file using that keyword, if it exists, or prompt you to do auto-tagging.

Auto tagging: It may take several minutes, depending on the amount of files, to manually categorise and search through all of the content stored on a smartphone's memory card before locating the desired file or piece of information. All of your tags will be safely archived in a database from which you may retrieve them at any time.

6. PROPOSAL OR DEVELOPMENT OF AN ALGORITHM

In order to categorise files based on the terms that appear in them, a powerful searching algorithm is utilised. The IEEE International Conference on Data Mining (ICDM) named C4.5, k-Means, SVM, Apriority, EM, PageRank, adaBoost, kNN, Naive Bayes, and CART as the top 10 data mining algorithms in December 2006. This is the algorithm we derived:

6.1. ALGORITHM

1. You may use a hash table to find certain terms by using the words as keys and setting the values to 0.

2. Repeatedly loop over the text and examine whether each word is a key in the hash table; if yes, add one to the corresponding value.

3. Using the matched words as keys and the counts as values, iterate through the hash table until you reach one of the non-zero entries.

Operates in O(N+M), where N and M are the number of words searched for and searched through, respectively.

6.2. ALGORITHM IMPLEMENTATION

1. Do a read-only open of the file.

2. Remove common terms (such as is, are, am, the, then, has, had, had, will, shall, and or, etc.) and conjunctions before reading the Start-Of-File (SOF) text.

3. When the following criteria are met, the keywords will be saved in the dictionary until the end of the file is reached:
- The value is increased by 1 if the key already exists.
- Set value to 0 if key does not exist.

4. Keys with the top 10 values should be listed.

5. Put these tags in the project repository to secure the keys.

6. Note the location of the important files and save them.

7. MODEL USED WORK FLOW

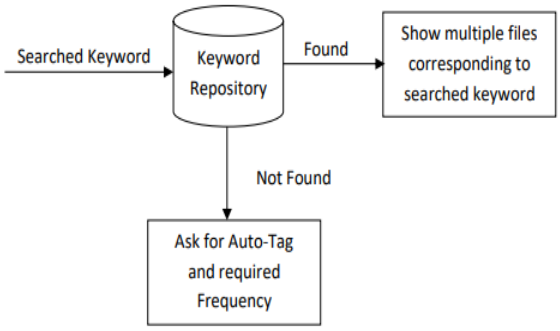

Figure-2: Tag Search Process

When manually tagging several files, the procedure must be repeated.

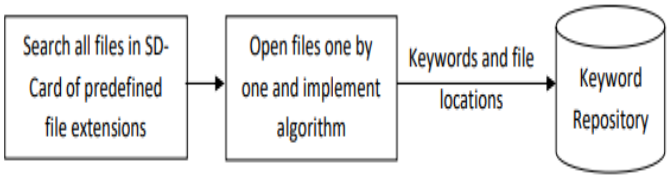

Figure-3: Auto-tag Process

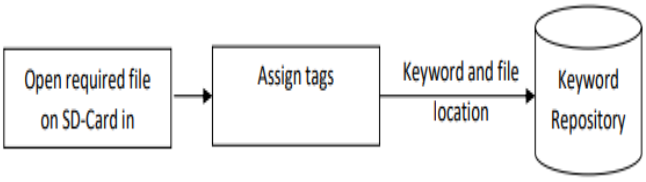

Figure-4: Manual-tag Process

8. RESULT AND ANALYSIS

The overarching purpose of this research endeavour is to provide strategies for the design of semantic mobile services that are more effective in terms of resource utilisation. In addition to this, the goal of the thesis is to present a categorization system that is meant to be complete from a theoretical standpoint with respect to the numerous composing techniques. Specifically, the classification system will focus on the many ways in which music may be composed. In addition to the numerous models that are a synthesis of that framework, the thesis gives a framework for the creation of tag-based smart content services. This framework may be found in the thesis. In this section of the essay, we are going to make an effort to search for heterogeneous contents that have been stored on the SD card of the mobile device but are not discoverable owing to the extensive range of file formats and content kinds. We are now working on the development of a repository that will be based on tags and will enable users to search for any phrase in order to identify the file (or files) that contain that word. This feature will be made possible thanks to the fact that we are currently working on the creation of a repository that will be based on tags. Because there are so many various types of files, it is challenging to acquire the result that we want because we need to build distinct code for each extension. This makes getting the result that we want more complex. This thesis provides an implementation for a Smart Content Tag Repository and focuses a substantial amount of attention on Android mobile devices as its primary research topic. Additionally, a number of other major contributions are included in the thesis. In addition to this, we have offered a comparative analysis of a variety of additional intelligent content aggregation and navigation algorithms available on different platforms. Lastly, the thesis provides evidence that this technology has been successfully implemented on the Android operating system, which is now among the most widely used operating systems for smart phones.

9. CONCLUSIONS

Plug-ins for other document types, document sources (RSS feeds, websites, e-mail, etc.), and language analyzers may be readily included into the SCAN platform. Extensions to the user interface may offer whole new domains of functionality. For instance, you may add a calendar plugin to your repository to see files in chronological order of their creation. It's possible that some other text analysis software is looking for related papers to a given one (search by pattern). What recent blog posts provide a glimpse into the research being conducted in this area, as well as how the Content Technologies field fits in with other areas of study? Standards, integration, content transfer, search, open source, and consumer-relevant technologies are only few of the areas covered. With Cloud material Management, current ECM systems may be supplemented with a low-cost, supplementary layer that improves secure collaboration amongst employees around live material. The capability of Cloud material Management may eventually grow to include support for capturing and archiving material as well. For some years, services like Google Docs and Zoho Docs have provided cloud-based document creation tools that may be used at the beginning of the content creation process. Recently, though, tools have emerged (like Google Wave) that make it possible for teams of writers to work in real time on a same manuscript.

REFERENCE

[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,