University of Ibadan
Journal of Science and Logics in ICT Research

SOLVIZ: A Document Visualization System Based On Topic Modelling

Odumuyiwa, V. T., Odujoko D. and Dalley, B.
Department of Computer Sciences, University of Lagos, Akoka, Nigeria

vodumuyiwa@unilag.edu.ng, odujoko.dml30@gmail.com, babsdalley@yahoo.co.uk

Abstract

The digital age has led to a voracious appetite for data as the world is gradually becoming a global village. Technology giants, businesses, schools and organization harness data in order to run their daily activities and beat competitors. Data has now become the crude oil of this global village. However, data could be structured, semi structured or unstructured. Understanding contents and features in the stack of data and getting hidden information stack could be very tedious and time consuming. Fortunately, a lot of information visualization techniques have been implemented to bring out meaning from the various forms of data. Meaningful information can be derived from data with the applications of these techniques. However, the application of some of these techniques pose some level of difficulties for the users. The presentation provided by some of these visualization techniques has rendered little help while leaving the user confused on getting around the information needed. This paper presents an information visualization abstraction for document visualization using the solar system. The abstraction led to the development of a document visualization tool tagged Solviz that provides an effective and powerful discovery of information in documents using topic modelling for extracting thematic elements from documents. It enables a user to explore hidden information in documents using a three dimensional space. Users can drill down into the document from topics to keywords to paragraphs in the document. The implementation also provides a data format which would allow anyone make use of this system efficiently.

Keywords: Document Visualization, Information Visualization, Topic Modelling, Multi-Document Visualization, Solviz.

1. INTRODUCTION

The amount of data in the world was estimated as 79 zettabytes in 2021. This is projected to rise to 181 zettabytes by 2025, including a significant increase in the number of documents, such as newspaper, articles, journals, books, criminal reports, medical reports, research papers, thesis, and so on. In addition, since the advent of Web 2.0 (Participatory Web), there has been an astronomical growth in the number of documents available on the World Wide Web. Many knowledge-rich contents are being generated on daily basis in all domains by both experts and non-experts. Web 2.0 has changed the ways in which information is collected and diffused. The users, who are meant to be consumers of information, are found to be playing the role of producers and consumers at the same time. The interactive nature of Web 2.0 has also broken the wall of partition separating experts from non-experts. Many online educational materials are being deployed on daily basis and interactions of learners around such materials are regularly being capitalized. It is estimated that the rate of data produced per day is 2.5 Quintillion bytes. This has led to so many research and career paths such as social media analytics, sentiment analysis, opinion mining, digital marketing, big data analytics and virtual learning. Based on this growth in information production, it is not uncommon today to hear people complain about information overload [1].

With such an enormous amount of data being produced, seeking manually for relevant information is synonymous to searching for a needle in a haystack. It is a very tedious task. With over 4.65 billion people using the internet, the quest for information could not be

Odumuyiwa, V. T., Odujoko D. and Dalley, B.T. (2022). SOLVIZ: A Document Visualization System Based On Topic Modelling, *University of Ibadan Journal of Science and Logics in ICT Research (UIJSLICTR)*, Vol. 7 No. 2, pp. 55 - 66

overemphasized. About half of the world's population are constantly seeking information on how to solve problems in their businesses, lives, home and community. With so much problems to solve and an urgent demand for solutions, very few people have the time and resources to go through all the data available in order to acquire meaningful and useful information.

Decision makers in organizations have so many documents to consult but with a limited time at their disposal to make necessary decisions which explain why there are always requests for executive summary. The same problem of information overload and limited time for learning applies to students at different level of educational pursuit. With Web 2.0, one can actually find learning contents for all domains online but having learning contents at one's disposal is different from extracting learning objects from those contents.

It is therefore imperative to develop methods, techniques and tools that can discover knowledge patterns in a collection of documents and automatically create graphical visualizations of the knowledge patterns in a way that can increase understanding, learning, analysis and interpretation of concepts and learning objects present in a document or collection of documents. Not only that, relationships between different concepts in different contexts can be visualized which can lead to innovation as it has been observed that when concepts cross domains, innovation becomes evident. Visualization involves techniques for creating images, illustrations, or animations to communicate a message [2].

Visualization has been adopted since the Paleolithic period, known as the Paleolithic art. The two main forms known to modern scholars are: "small sculptures and monumental paintings, incised designs, and relief on the walls of caves". This implies that visualization has always been the most efficient way to communicate a message [3]. Information visualization provides various graphical components/menus and user interface for exploration and manipulation of large numbers of items, extracted from far larger datasets [4]. It helps the user take decisions, and discover patterns in their data which was not known before. It enables the user give clearer explanation on the content of their data.

There are several challenges in Information visualization. One challenge is providing the right data format for the visualization tool. For the visualization tool to work properly, the data format must be presented in a manner easy for the tool to process. Another problem encountered in information visualization is the ability to find related information. This paper presents a solar system abstraction for the development of a multi-document visualization tool name "Solviz" and the JSON data format for data presentation to the visualization tool.

The remainder of this paper is presented as follows: section two presents a review of the literature, section three discusses the methodology for the research by detailing the solar system abstraction and presenting the algorithm for topic modelling and procedure for information retrieval, section four presents the results and section 5 concludes the paper.

2. LITERATURE REVIEW

2.1 Introduction to Document Visualization

Everything is document and document is everything. This looks provocative at the first look but giving it a second thought, it begins to make some sense. The earth is filled with objects and every object can be represented as a document [5] by generating a list of attributes and values that describes the object. An individual can be represented by his/her profile which is expressed inform of a textual document. The content of an audio-visual object can be represented by a set of metadata and also by a brief textual description of the object. In the digital world, automatic processing of textual documents has gained a lot of attention over the years and great advancements have been made in this area. Leveraging on the advancements made in automatic document processing especially as it relates to text mining to enhance learning experience of learners will not but be a welcomed idea.

Learners are not necessarily students in the traditional/conventional sense of it. In fact, anytime an individual notices he/she needs information to solve a problem, take a decision, or do whatever, such a person at that moment in time possesses an anomalous knowledge state [6] [7] [8]. An additional knowledge is needed by the person in order to resolve the anomaly in his/her knowledge state at that point in time. The process

of acquiring this additional knowledge is known as learning. It means that a student can learn to master a new knowledge area, a manager can learn in order to make right decisions, even a family intending to go for holiday can learn in order to make a right decision of a holiday destination to go to. The learning process can take from few seconds to years depending on the kind of problem to solve. The learning period is not necessarily a function of lack of information but it could be related to inability to pick a relevant information source out of a plethora of sources available. Even when a right information source is picked, it could be related to inability to extract relevant learning object from the information source.

The ability of the computer as a fast machine that processes large set of data in a very short time can come into play in enhancing interactions around learning documents and ensuring that learners do not spend endless time in extracting learning objects from document. The paper therefore aims at enhancing learning at all levels by developing single and multiple document visualization methods for concept and learning objects extraction.

According to Schneiderman & Plaisant [4], Information visualization is defined as "the use of interactive visual representations of abstract data to amplify cognition". They explained further that "information visualization provides compact graphical representations and user interfaces for interactively manipulating large numbers of items (10^2 to 10^6), possibly extracted from large datasets". Information visualization is a large field in computer sciences with a lot of research work and subfields. Common subfields include: text visualization [9], high dimensional data visualization, social network visualization and document visualization.

Text visualization [10] in simple terms focuses mainly on the text level without considering the attributes and metadata. High dimensional data visualization focuses on visualizing multidimensional data. Social network visualization focuses on visualizing the relationship among people in form a node-link data. Document visualization focuses on visualizing textual document information while concentrating on the attributes and metadata of the documents.

This main interest of this paper is on document visualization. Document visualization can be defined as "a class of information visualization techniques that transforms textual information such as words, sentences, documents, and their relationships into a visual form enabling users to better understand textual documents and to lessen their mental workload when faced with a substantial quantity of available textual documents" (Qihong et al, 2014). Document visualization enables easy identification of the key themes discussed in a document and reduces the time spent by users in exploring a document [11].

Ben Shneiderman proposed a general principle for document visualization indicating seven basic data types that can be visualized and seven basic tasks that visualization tools should support. This section discusses the seven data types (1D, 2D, 3D, Multidimensional, Tree, temporal and Network), different visualization tasks and presents some existing visualization tools highlighting the data types they support and the visualization tasks possible in such tools.

2.1.1 Data Types

The seven basic data types accompanied with the general principle mantra are useful to describe the visualization that have been developed and to characterize the classes of problems that users encounter.
1D Linear data types are one-dimensional data or lists of data items organized by a single feature. 2D Map Data, also referred to as planar data, includes geographic maps, floorplans, and newspaper layout. 3D World Data represents three-dimensional relationships among real world objects. Multidimensional Data are data in which items with n attributes become points in an n-dimensional space.

Tree Data shows relationship in form of hierarchies. Items (nodes) in the tree except the root node have at least one parent node. Items also points to chid node(s) except the leave nodes. Tree structure does not allow formation of cycles in the relationship among nodes. Temporal data captures temporal context relating to data observed on item of interest and such data varies over time [12] [13]. Network Data captures relationship among items that are not supported in a tree structure. These seven data types provides an abstraction of the various

possibilities of data presentation in reality. However it is worth noting that there are several variants of these seven types.

2.1.2 Visualization Tasks

After processing the data and getting all necessary information needed, it is very important to consider the manner in which this data will be displayed to the user. The manner in which the information is displayed should allow the user gain insight into the original data, perform some exploration task and reduce stress in acquiring this information. Schneiderman & Plaisant propose a taxonomy comprising of seven tasks which include: overview, zoom, filter, and details-on-demand, relate, history, and extract [4].

Overview Task: with so much information available to display to the users, displaying it all at once can be overwhelming. This could be slightly better than the raw data, but it still does not give the user the flexibility needed for exploration. Overview is basically a brief summary of the whole data.

Zoom Task: With a good overview provided, the user might also want to focus on a particular area of interest. Zooming is a way to achieve this. Tools should be provided to enable them to control the zoom focus and the zoom factor.

Filter Task: While zooming deals with focusing on an area of interest, filtering deals with removal of unwanted or irrelevant items. Filtering gives users control over the contents displayed and enables them to focus on their interests.

Details-On-Demand Task: Oftentimes, users will need to get more information about an item. This task makes it possible for users to select an item and view more details about the item.

Relate Task: entails enabling users to view relationships among items based on different parameters such as proximity or by containment using different colour coding and connecting lines.

History Task: this task enables users to view the history of actions performed hence facilitating reuse, redo, replay and refinement of such past actions.

Extract Task: with this task, users are provided with functionalities to extract either dataset or the displayed visualization and save, copy or transfer such for use elsewhere.

2.1.3 Document Visualization Methods

There are three common document visualization categories: single document visualization, multi-document visualization and extended document visualization. Single document visualization is mainly concerned with the words and contents of a particular document [14]. Examples include Tag Clouds, Wordle, Word Clouds [15]. Multi-document visualization deals with concepts and relationships among two or more documents and Extended document visualization concentrates more on comprehensive tasks and involves other attributes beyond the content of the documents. It is often applied in specific fields such as social media and search.

2.1.4 Existing Systems

Name	**Description**	**Data Type**	**Task**	**Visualization Method**
Wordle	Wordle is a visualization tool that displays texts in difeerent sizes based on their frequency of occurrence in a given document [16].	One dimensional linear data	Overview and extract task.	Single document visualization
TextArc	TextArc visualizes documents by showing pictorially word frequency and distribution as well as association [17].	One dimensional linear data	Overview task, detail-on-demand task, filter task and relate task.	Single document visualization

DocuBurst	Creates a visualization indicative of the lexical and semantic content of a document using a combination of word frequency and structure in lexical databases [18].	One dimensiona l linear data	supports all tasks except the history and extract task	Multi-document visualization
Phrase nets [19]	Phrase nets shows words from unstructured document in the form of a graph and using edges between words to show user defined relationship between them. Frequency is reflected in the size of the words while the flow and strength of the connection between two words are reflected in the direction and width of the arrow respectively.	One dimensiona l linear data	Overview, Zooming, Filtering task, and Relate task.	Single document visualization
ORCAESTRA	ORCAESTRA is a system for news comments organization and visualization. ORCAESTRA is an acronym for **OR**ganizing news **C**omments using **A**spects, **E**ntity and **S**entiment ex**TRA**ction [20].	One dimensiona l linear data and network data	Overview tasks Zoom task, Filtering task Details-on-demand, relate task.	Extended document visualization
Digital Attack Map	It displays global Distributed Denial of Service (DDoS) activity on any given day [21]. Dotted lines are used to show the attacks from source to destination countries.	Multidimen s-ional and network data	Overview task, Zoom task. Filtering task Details-on-demand, relate task. Extract task	Extended Document visualization.
ThemeRiver	Uses a river flow metaphor to visualize thematic changes over time in a set of documents [22]. It is useful for identifying how themes evolve as well as time-related patterns and relations in a large corpora.	Temporal data	Overview and relate task	Multi document visualization

2.2 Document Visualization using Topic Modelling

Topic modelling has been widely applied in automatically inferring the thematic focus of documents especially in information retrieval systems. In this context of this paper, topic modelling using Latent Dirichilet Allocation (LDA) algorithm is applied to extract topics (themes) that describe the content of a document. Latent Dirichlet Allocation (LDA) is a generative statistical approach to topic modelling which was proposed by David Blei and others in 2003 [23]. It expresses the semantic content of a corpus in a concise manner hence enabling one to compare how similar one document is with another by looking at how similar the corresponding topic mixtures are. LDA model incorporates a Dirichlet (A multivariate probability distribution) prior to its topic distribution. This means it is a probability distribution that estimates a quantity before any evidence is taken into account. LDA also has two classes of variables, that is, the unobserved and the observed variables. The unobserved variables here are the topics and the observed variables are words. The number of topics to be inferred must be specified by the user beforehand. The goal of LDA is to infer the hidden variables - topics - conditioned on the observed variables. In order to do this, it assumes that a topic is a probability distribution over words; a document is a random mixture over topics; and a word is drawn from one of these topics.

3. METHODOLOGY

3.1 Visualization Abstraction

The visualization of documents using the solar system could be puzzling if the concept of the solar system is not understood first. This section seeks to clarify the various concepts adopted from the solar system.

According to the National Aeronautics and Space Administration (NASA), a solar system is a star and all the objects that travel around it. These objects include: planets, moons, asteroids, comets and meteoroids. Most stars host their own planets. There can be more than one star in a solar system. A solar system with two stars is known as a binary star system. It is known as a multi-star system if it has three or more stars. Our solar system consist of the sun (our star) and objects revolving around it. These objects include: 8 planets (Mercury, Venus, Earth, Mars, Jupiter, Saturn, Uranus, Neptune), 5 dwarf planets (Pluto being the popular one); 173 moons (with 149 known and 24 provisional), more than 3400 comets, and more than 715, 000 asteroids.

As earlier discussed, an information visualization system needs to adopt seven basic task: overview, zoom, filter, details-on-demand, relate, history, and extract. The goal of this section is to give a breakdown of how the solar system was used to achieve these tasks.

The data format given in section 3.2 shows that the solar based visualization method can be used for both single and multiple documents. Since the document is the core component, a star or groups of stars is used to represent it. For a single document, a star is used while two or more stars are used for multiple documents. The star used in this case is the sun. The next component are topics. Topics are represented using the eight planets and a dwarf planet (Pluto). Naturally, planets come in different sizes with varying distance from the sun. The distance of a planet from the sun determines the amount of sunshine it gets. Rather than making use of planet with different sizes, we rely on its closeness to the sun to show similarity to the document. The closeness is determined by the rank of the topic.

The last component of the data format are keywords. Planets have landscapes, mountains and valleys. Each planet also holds the keywords it contains. The seven tasks were achieved using this concept as follows:

- **Overview**: A group of stars (sun) is used to show the documents. This gives the user the amount of documents being visualized. The planets also give an overview of all topics contained in a document. Finally, each planet gives an overview of all keywords present in a topic.
- **Details-on-demand**: The first overview the user encounters is the group of stars. When a star is clicked, the solar system is displayed. When a planet is clicked, the planet gains focus and its keywords are displayed. When a keyword is clicked, the document content is then retrieved.
- **Relate**: Relationship is shown among topics from the solar system. The closer a planet is to the sun, the more its relevance.
- **Zoom**: Users can zoom in and out using the mouse scroll wheel or trackpad.
- **Filter**: The visualization system has its own information retrieval system (IRS). The essence of its IRS is to provide room for possible customization.
- **History**: users' actions are kept while users navigate across various actions.
- **Extract**: The only extraction process supported is copying the text.

Figure 1 gives a conceptual view of the proposed system. It should be noted however that the system did not try to achieve a strict similarity to the actual solar system. It indeed got close and could be further improved to be very similar.

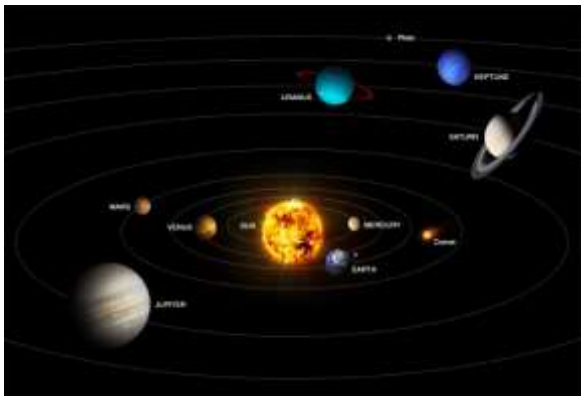

Figure 1: Solar System Illustration (*http://chandra.harvard.edu/graphics/resources/illustrations/solsys/sol_sys_illD.jpg*)

3.2 Data Format for Solviz

The data format is basically the structure of the data that the visualization system can interpret in order to produce the right result. This was done in order to make the system flexible and adaptable to any textual document.

The data format specified for the system is as follows:

```
[
  {
    "document": "doc1",
    "topics":[
      {"label": "topic1", "rank": 10},
      {"label": "topic2", "rank": 5},
      {"label": "topic3", "rank": 3}
    ],
    "keywords":[
      {"label": "word1", "rank": 40},
      {"label": "word2", "rank": 59},
      {"label": "word3", "rank": 20}
    ]
  }
]
```

Figure2: Data format for the Solviz visualization system

The data format is a JavaScript Object Notation (JSON). This is the common format for data representation on the web. The components of the format specified include:

- **document**: This represents the document name.
- **topics**: This is an array of topic object. A topic can have a *label* and *rank*. The rank of the topic is used to determine the most prominent topic in the document. The label is synonymous to the topic or name.
- **keywords**: This is also an array of keyword object. A keyword can have a *label* and *rank*. The rank of a keyword is used to determine the most prominent keyword around a topic.

The data format itself is an array of document object. This simply means, we can have more than one document in the data. With this, the system can be used for both single document and multi-document visualization.

3.3 LDA algorithm implementation for topic extraction

There are two parts to the algorithm. The first part is the initialization phase and the second part is the inference phase, which makes use of Collapsed Gibbs Sampling [24]. The two parts as implemented in this work are presented in the flow charts in Figures 3 and 4. LDA only gives a distribution of words over a topic. It does not provide any label for the topic. In the context of this work, topic labels are important for users to have an overview of the main focus (or themes) of a document.

In order to derive labels for the topics, SPARQL query language was used to query the dbpedia ontology to extract two subjects for each word in a topic distribution. The list of subjects created for each topic was compared with the words in the original topic distribution using cortical retinasdk library to find the semantic similarities and the best subject semantically resulting from the comparison was picked as the label for the topic.

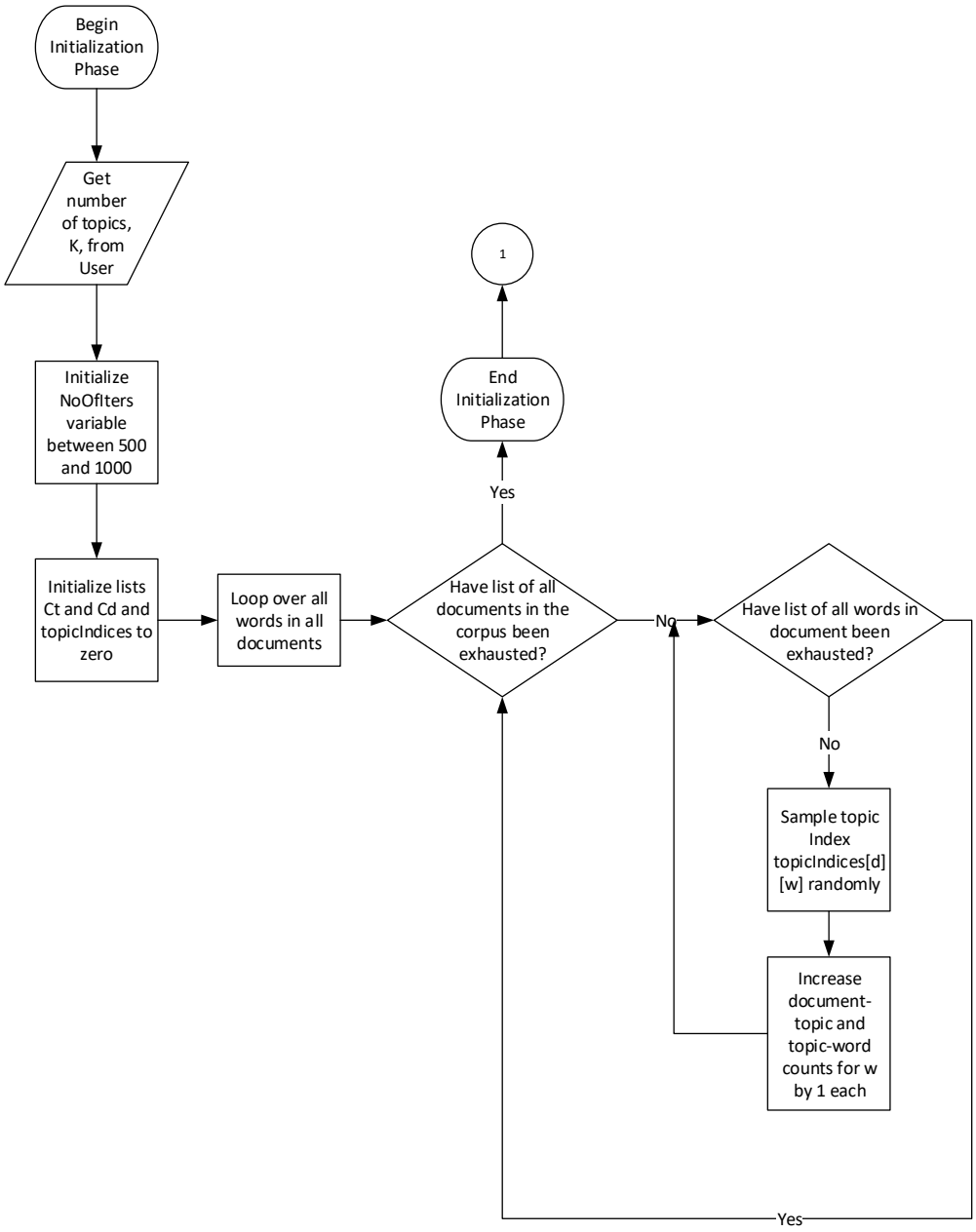

Figure3: Flowchart of the initialisation phase of the LDA algorithm

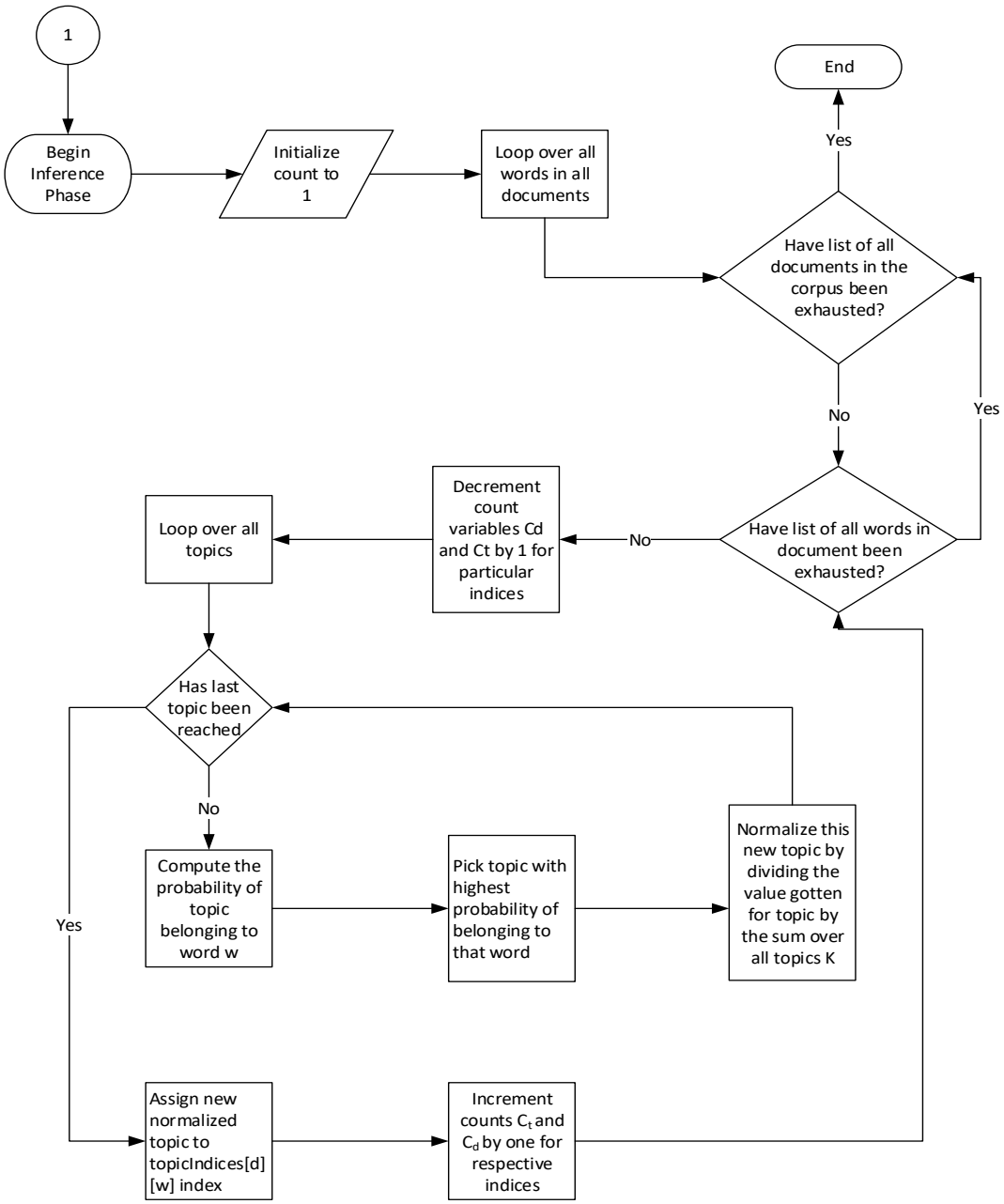

Figure 4: Flowchart of the inference phase of the LDA algorithm using the collapsed Gibbs Sampling

3.4 Information Retrieval for the "filter" task

Information retrieval deals with getting important materials from a large set of documents in order to satisfy an information need. The aim of visualization is to enable users get useful information within a short period of time [25]. Therefore, this is a supporting feature in order to make the solar system based technique robust and more useful. The methodology for the information retrieval system as shown in Figure 5 can be described as follows:

Preprocessing: Before the documents could be used for the information retrieval system, tokenization was done, followed by stemming (removal of tokens derivational affixes), part of speech tagging (identifying and categorizing tokens with similar grammatical properties, lemmatization, and removal of stop words. Likewise queries are also preprocessed before they are matched against the document using cosine similarity.

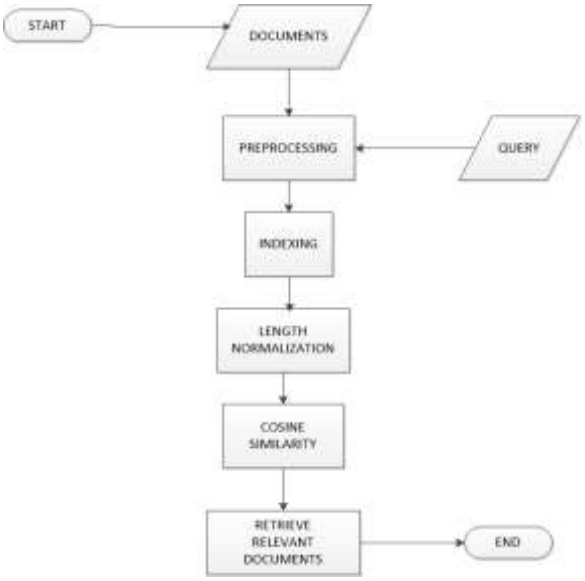

Figure 5: Flowchart for the information retrieval component of the Solviz system

Indexing: An index is a data structure which is used for representing all documents of an information retrieval system. An index basically contains the terms and the document identifier. Usually the two parts of the index are dictionary (consisting of all terms) and postings (consisting the list of the documents containing the term and may also include the term's position). In the proposed system, we made use of a positional index. A positional index stores the positions of the terms as they appear.

Ranking: This involves returning documents based on their relevance with the user's query. The Vector Space Model was used. The Vector Space model is used to represent document as vectors in other to rank them.

4. RESULTS AND DISCUSSION

The Solviz system was implemented using Python programming language for the backend and Javascript for the frontend. As discussed in section 2, the LDA algorithm was used to obtain the probability distribution over topics which represents the thematic structure of documents to be visualized. Figure 6 shows a single document visualization on Solviz. The sun which is labelled as "body" is the document being visualized and the eight planets are the topics extracted from the documents. The distance of the planets from the sun depicts the degree of similarity of the topics to the content of the document which implies that the closer the planet (topic) is to the sun (document) the more relevant the topic is in describing the content of the document. Because we are using the solar system abstraction for visualization, the planets (topics) revolve around the sun (document).

Figures 7, 8, 9 and 10 show the West, North, East and South views of the visualization as the planets (topics) revolve around the sun (document). The topics are also displayed on the left pane as shown in Figure 6. Clicking on a topic gives the keywords that form the topic distribution and clicking on the keywords launches a query on the information retrieval system module of Solviz to retrieve relevant document content.

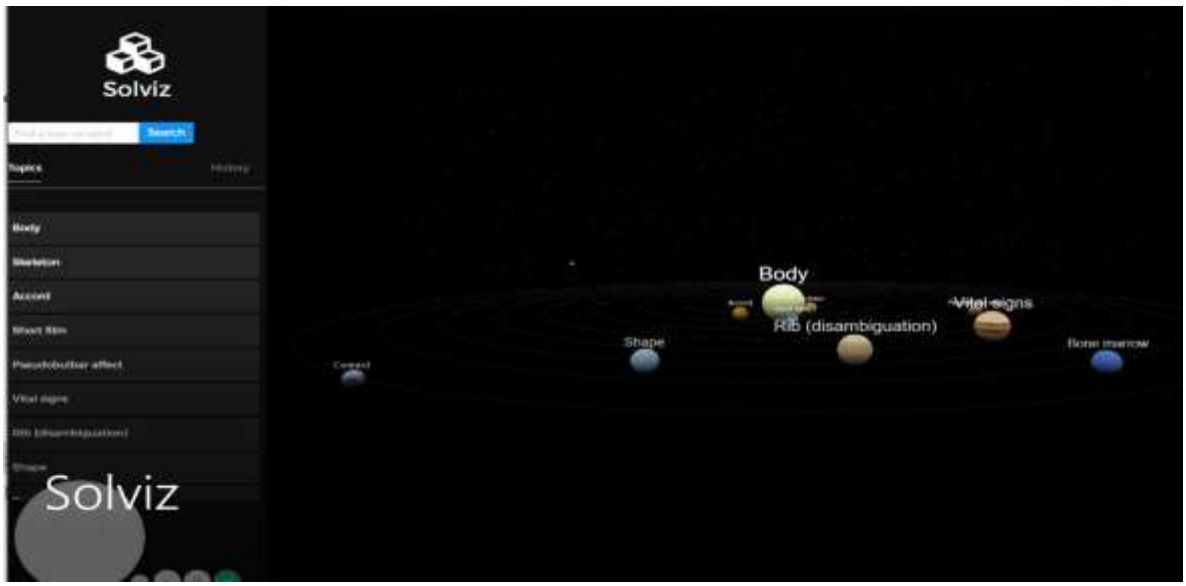

Figure 6: Solviz visualization system showing single document visualization

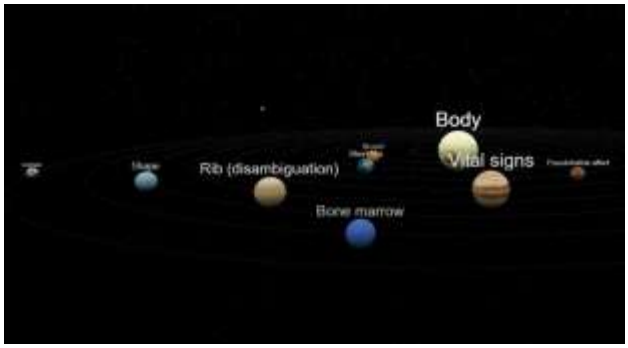

Figure 7: West view of the visulaization as the planets (topics) revolve round the sun (document)

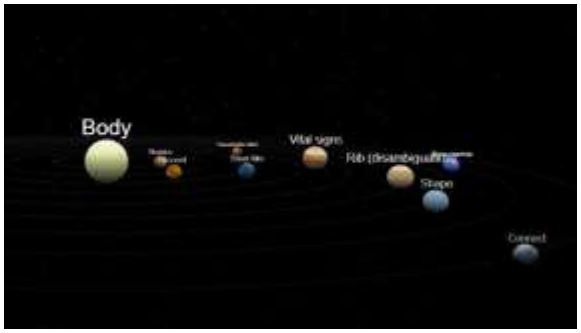

Figure 8: North view of the visulaization as the planets (topics) revolve round the sun (document)

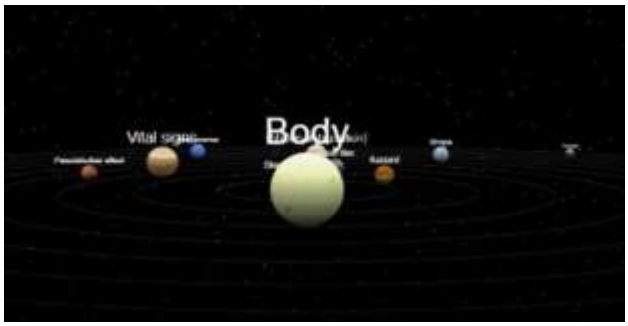

Figure 9 : East view of the visulaization as the planets (topics) revolve round the sun (document)

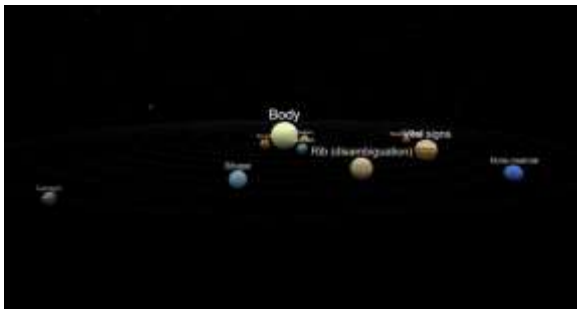

Figure 10: South view of the visulaization as the planets (topics) revolve round the sun (document)

5. CONCLUSION

Information visualization especially document visualization enhances quick identification, presentation and understanding of the core content of a document. Topic modelling enables the machine to infer automatically the topics of interest in a given document or collection of documents. This paper discussed a visualization abstraction for multi-document visualization based on the solar system. The abstraction and the data format created enable the development of Solviz for document visualization. Solviz takes in a multidimensional data and visualizes the data while supporting the seven task taxonomy of an effective visualization tool. The results obtained show that Solviz can carry out single document visualization effectively. Further work is needed to ensure effective multi-document visualization implementation on Solviz based on the multi-document visualization abstraction documented in this paper. Future work will also explore or create other algorithms for extracting relevant topics from documents.

References

[1] D. A. Keim, F. Mansmann, J. Schneidewind and H. Ziegler, "Challenges in visual data analysis," in *Tenth International Conference on Information Visualisation (IV'06),*, 2006.

[2] M. O. Ward, G. Grinstein and D. Keim, Interactive data visualization: foundations, techniques, and applications, Natick, MA 2010: AK Peters, Ltd, 2010.

[3] D. Dharmayanti, A. M. Bachtiar and F. N. Fakhrul, "Data Visualization For Content Marketing Domain In Social Media," *Journal of Engineering Science and Technolog*, vol. 16, no. 1, pp. 339-349, 2021.

[4] B. Shneiderman and C. Plaisant, Designing the user interface: Strategies for effective human-computer interaction, India: Pearson Education, 2010.

[5] M. K. Buckland, "What Is a ‘‘Document’’?," *Journal of the American society for information science*, vol. 48, no. 9, pp. 804-809, 1997.

[6] N. J. Belkin, "Information concepts for information science," *Journal of Documentation*, vol. 34, no. 1, pp. 55-85, 1978.

[7] V. Odumuyiwa, "Exploiting User Generated Content during Explicit Collaboration for Knowledge Discovery," Lille, 2011.

[8] A. David, V. Odumuyiwa, T. Oguntunde and O. Akerele, "ICT-Based Environment for Training: Trends, Opportunities and Challenges," *Communication, technologie et développement David A., Odumuyiwa V., Oguntunde T. and Akerele O.*, vol. 1, no. 1, pp. 34-46, 2014.

[9] G. R. Choudhary and I. Sharma, "Open source text visualization tools: A comparative analysis," *JIMS8I International Journal of Information Communication and Computing Technology*, vol. 9, no. 2, pp. 521-528, 2021.

[10] K. Kucher and A. Kerren, "Text visualization techniques: Taxonomy, visual survey, and community insights," in *Visualization Symposium (PacificVis)*, 2015.

[11] J. K. Chou and C. K. Yang, "PaperVis: Literature review made easy," in *Computer Graphics Forum*, Oxford, UK, 2011.

[12] C. S. Jensen and R. T. Snodgrass, "Temporal data management," *IEEE Transactions on knowledge and data engineering*, vol. 11, no. 1, pp. 36-44., 1999.

[13] C. Daassi, L. Nigay and M.-C. Fauvet, "Visualization Process of Temporal Data," *Lecture Notes in Computer Science*, pp. Chaouki Daassi, Laurence Nigay, and Marie-Christine Fauvet., 2004.

[14] G. Qihong, Z. Min, L. Mingzhao, L. Ting, C. Yu and Z. Baoyao, "Document visualization: an overview of current research.," *Wiley Periodicals*, vol. 6, no. 52. Qihong G., Min Z., Mingzhao. L., Ting. L. ,Yu. C., Baoyao. Z., pp. 19-35, 2014.

[15] V. M. Rajan and A. Ramanujan, "Architecture of a Semantic WordCloud Visualization," in *Second International Conference on Networks and Advances in Computational Technologies*, 2021.

[16] F. B. Viegas, M. Wattenberg and J. Feinberg, "Participatory visualization with Wordle," *IEEE transactions on visualization and computer graphics*, vol. 15, no. 6, pp. 1137-1144., 2009.

[17] W. B. Paley, "Textarc: Showing word frequency and distribution in text," in *IEEE Symposium on Information Visualization*, 2002.

[18] C. Collins, S. Carpendale and G. Penn, "Docuburst: Visualizing document content using language structure," *Computer graphics forum*, vol. 28, no. 3, pp. 1039-1046, 2009.

[19] F. Van Ham, M. Wattenberg and F. B. Viégas, "Mapping text with phrase nets," *IEEE transactions on visualization and computer graphics*, vol. 15, no. 6, pp. 1169-1176, 2009.

[20] R. E. Prasojo, F. Darari and M. Kacimi, "ORCAESTRA: Organizing news comments using aspect, entity and sentiment extraction," in *IEEE VIS*, 2015.

[21] M. Baykara, U. Gurturk and R. Das, "An overview of monitoring tools for real-time cyber-attacks," in *6th International Symposium on Digital Forensic and Security (ISDFS)*, 2018.

[22] S. Havre, E. Hetzler, P. Whitney and L. Nowell, "Themeriver: Visualizing thematic changes in large document collections," *IEEE transactions on visualization and computer graphics*, vol. 8, no. 1, pp. 9-20, 2002.

[23] D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, pp. 993-1022, 2003.

[24] I. Porteous, D. Newman, A. Ihler, A. Asuncion, P. Smyth and M. Welling, "Fast collapsed gibbs sampling for latent dirichlet allocation. In," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*.

[25] M. Hearst, "User interfaces and visualization," in *Modem Information Retrieval*, R. Baeza-Yates and B. Ribeiro-Neto, Eds., Reading, MA, Addison-Wesley, 1999.