IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 7, July 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 7, July 2024

|DOI: 10.15680/IJIRSET.2024.1307090|

Data Mining with Context Remembering: Harnessing the Power of Disparate Data Sources

Bhupendrasinh Thakre

Walmart, USA

ABSTRACT: Data mining with context remembering has emerged as a powerful technique for extracting meaningful insights from disparate data sources in the era of big data. This article explores the concepts and benefits of context-aware data mining, focusing on the techniques of cosine similarity and feature engineering. By leveraging approaches similar to those used in Natural Language Processing (NLP), data mining with context remembering derives reasoning from the contextual search and resemblance of data points, enabling businesses to unify and analyze heterogeneous data effectively. The article delves into the application of cosine similarity and feature engineering in identifying contextually similar items and overcoming differences in word choices and formats. The benefits of these advanced data mining techniques are discussed, including harnessing the full potential of data, deriving accurate and meaningful insights, and dealing with complex and varied information. As the volume and variety of data continue to grow, the importance of context-aware data mining will only increase, making it an essential tool for businesses seeking to maintain a competitive edge in the era of big data.

KEYWORDS: Data mining, Context remembering, Disparate data sources, Cosine similarity, Feature engineering

I.INTRODUCTION

In the era of big data, businesses are inundated with vast amounts of information originating from various sources, both internal and external. The ability to effectively harness and derive meaningful insights from this data has become a critical factor in maintaining a competitive edge [1]. However, the challenges associated with data mining have grown exponentially as the volume and variety of data continue to increase. One of the most significant hurdles is the presence

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 7, July 2024

|DOI: 10.15680/IJIRSET.2024.1307090|

of disparate data sources that lack a consistent structure or format, making it difficult to match and analyze the information effectively [2].

Data mining with context remembering has emerged as a powerful solution to this problem, enabling businesses to unify and extract valuable insights from heterogeneous data sources. By leveraging techniques similar to those used in Natural Language Processing (NLP), context-aware data mining derives reasoning from the contextual search and resemblance of data points [3]. This approach allows for the identification of similarities between data items, even when they are expressed using different words or formats, ultimately enabling businesses to harness the full potential of their data and derive accurate and meaningful insights.

In this article, we will explore the concepts of data mining with context remembering, focusing on the techniques of cosine similarity and feature engineering. We will delve into the benefits of these advanced data mining techniques and demonstrate how they can be applied to effectively handle disparate data sources, unlocking the hidden value within the complex and varied information collected by modern businesses.

II. DATA MINING WITH CONTEXT REMEMBERING

Data mining with context remembering is an advanced technique that enables the extraction of meaningful insights from disparate data sources by considering the contextual relationships between data points [4]. This approach goes beyond traditional data mining methods, which often struggle to handle heterogeneous data effectively [5].

1. Handling multiple data sources
Context-aware data mining allows businesses to unify and analyze data from various sources, such as databases, social media, and other resources, by identifying the contextual similarities between data items [6].

2. Matching data despite differing values with the same meaning
By considering the context in which data points appear, this technique can match and group similar items even when they are represented using different words, formats, or structures [7].

Data mining with context remembering leverages techniques similar to those used in NLP, such as semantic analysis and topic modeling, to understand the meaning and context of data points [8]. Context-aware data mining algorithms derive reasoning by searching for contextual similarities and resemblances between data items, enabling the identification of hidden patterns and relationships [9].

Aspect	Traditional Data Mining	Data Mining with Context Remembering
Data sources	Homogeneous	Heterogeneous
Data preprocessing	Limited	Extensive
Contextual information	Not considered	Incorporated
Similarity measures	Basic (e.g., Euclidean)	Advanced (e.g., cosine similarity)
Insight quality	Moderate	High

Table 1: Comparison of traditional data mining and data mining with context remembering [7-9]

III. COSINE SIMILARITY AND FEATURE ENGINEERING

Cosine similarity and feature engineering are two essential techniques used in advanced data mining, particularly when dealing with text data or high-dimensional datasets. These techniques enable data mining algorithms to effectively measure the similarity between data points and extract the most relevant features for analysis.

A. Cosine similarity
Cosine similarity is a metric used to measure the similarity between two non-zero vectors in a high-dimensional space [10]. In the context of data mining, cosine similarity is often used to determine the similarity between text documents, data points, or feature vectors. The metric calculates the cosine of the angle between two vectors, with the resulting

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 7, July 2024

|DOI: 10.15680/IJIRSET.2024.1307090|

value ranging from -1 to 1. A cosine similarity of 1 indicates that the vectors are identical, while a value of 0 suggests that the vectors are orthogonal or unrelated, and -1 indicates opposite context or meaning.
To calculate the cosine similarity between two vectors A and B, the dot product of the vectors is divided by the product of their Euclidean lengths or magnitudes [12]. The formula for cosine similarity is as follows:

$$\text{Cosine Similarity}(A, B) = (A \cdot B) / (||A|| * ||B||)$$

where $A \cdot B$ represents the dot product of vectors A and B, and $\|A\|$ and $\|B\|$ denote the Euclidean lengths of the respective vectors.

The dot product of two vectors $A = (a_1, a_2, ..., a_n)$ and $B = (b_1, b_2, ..., b_n)$ is defined as:

$A \cdot B = \sum_{i=1}^{n} (a_i * b_i) = a_1b_1 + a_2b_2 + ... + a_nb_n$

In this notation, Σ represents the summation symbol, i is the index variable, and n is the number of elements in each vector. The dot product is calculated by multiplying the corresponding elements of the two vectors and then summing the results.

The Euclidean length or magnitude of a vector $A = (a_1, a_2, ..., a_n)$ is defined as:

$\|A\| = \sqrt{(a_1^2 + a_2^2 + ... + a_n^2)}$

In data mining, cosine similarity is particularly useful for comparing text documents, as it is invariant to document length and focuses on the proportions of the terms rather than their absolute frequencies [21].

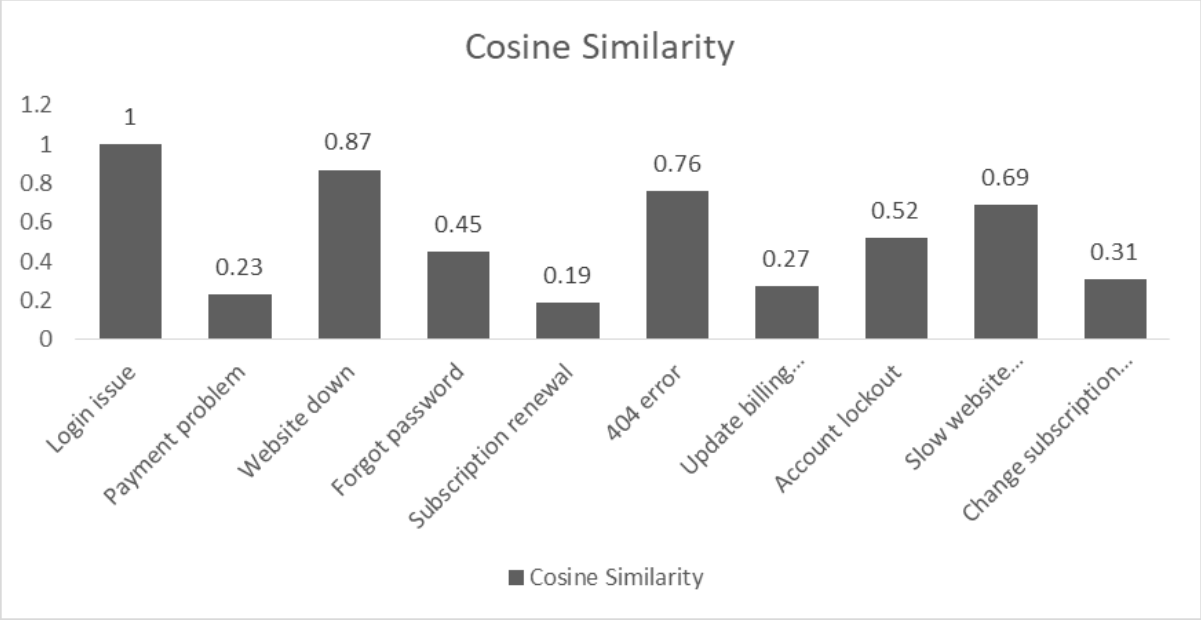

Figure 1: Sample dataset for customer support ticket similarity using cosine similarity [29]

B. Feature engineering
To apply cosine similarity or other similarity measures in data mining, text data must first be converted into numerical vectors. This process is known as text vectorization or feature extraction. Common techniques for converting text data into vectors include:

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 7, July 2024

|DOI: 10.15680/IJIRSET.2024.1307090|

a. Bag-of-words: This approach represents a text document as a set of its words, disregarding grammar and word order. Each unique word in the corpus becomes a feature, and the feature values are the frequencies or occurrences of the words in the document [13].

b. TF-IDF (Term Frequency-Inverse Document Frequency): TF-IDF is an extension of the bag-of-words model that reflects the importance of a word in a document relative to the entire corpus. It considers both the frequency of a word within a document (term frequency) and the inverse of the frequency of the word across all documents (inverse document frequency) [22].

c. Word embeddings: Word embeddings are dense vector representations of words that capture semantic and syntactic relationships. Techniques like Word2Vec [23] and GloVe [24] learn these embeddings from large text corpora, enabling words with similar meanings to have similar vector representations.
Feature engineering involves selecting, creating, and transforming features to improve the performance of data mining algorithms [14]. In the context of text data, feature engineering techniques may include:

a. **Removing stop words:** Stop words are common words (e.g., "the," "and," "is") that often carry little meaning and can be removed to reduce the dimensionality of the feature space [25].
Here's an example of removing stop words from the sentence "John Doe visited St. Mary's Hospital on 2023-04-15 with a fever.":

Original sentence:
John Doe visited St. Mary's Hospital on 2023-04-15 with a fever.
Sentence with stop words removed:
John Doe visited St. Mary's Hospital 2023-04-15 fever.
In this example, we remove the following stop words from the sentence:

- "on"
- "with"
- "a"

Stop words are common words that frequently appear in text but often carry little meaningful information. These words, such as prepositions, conjunctions, and articles, can be removed from the text to reduce the dimensionality of the feature space and improve the efficiency of text mining tasks.

By removing stop words, we can focus on the more informative and relevant words in the text, such as "John Doe", "St. Mary's Hospital", "2023-04-15", and "fever". This can help in tasks like document classification, clustering, or similarity analysis, where the presence or absence of specific informative words can be more indicative of the content and meaning of the text.

Removing stop words can also help reduce the computational complexity of text mining algorithms, as there are fewer unique words to process. This can lead to faster processing times and more efficient use of computational resources.

It's important to note that the list of stop words can vary depending on the specific language, domain, and task. In some cases, it may be necessary to create a custom list of stop words that are specific to the problem at hand. Additionally, it's essential to be cautious when removing stop words, as in some cases, they may carry important information or context, particularly in idiomatic expressions or domain-specific terminology. Overall, removing stop words is a common text preprocessing technique in data mining and natural language processing that can help improve the efficiency and effectiveness of text analysis by focusing on the most informative and relevant words in the text.

b. **Stemming and lemmatization:** These techniques reduce words to their base or dictionary forms, helping to group related words and reduce the number of unique features [26].

Here's an example of stemming and lemmatization for the sentence "John Doe visited St. Mary's Hospital on 2023-04-15 with a fever.":

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 7, July 2024

|DOI: 10.15680/IJIRSET.2024.1307090|

Original sentence:
John Doe visited St. Mary's Hospital on 2023-04-15 with a fever.
Stemmed sentence:
John doe visit st. mary' hospital on 2023-04-15 with a fever.
Lemmatized sentence:
John Doe visited St. Mary's Hospital on 2023-04-15 with a fever.
In this example, we apply stemming and lemmatization techniques to the words in the sentence:

Stemming:
- "visited" is reduced to its stem "visit"
- "Mary's" is reduced to "mary'"
- "Hospital" is reduced to "hospital"

Stemming algorithms, such as the Porter stemmer, strip off word endings based on a set of predefined rules. This process may sometimes result in stems that are not actual words, like "hospital" for "Hospital".

Lemmatization:
- "visited" is reduced to its base form "visit"
Lemmatization, on the other hand, uses a vocabulary and morphological analysis to reduce words to their base or dictionary forms, known as lemmas. In this example, "visited" is lemmatized to "visit", which is the base form of the verb.

By applying stemming and lemmatization, we can group related words and reduce the number of unique features in the text data. This can help improve the efficiency and accuracy of text mining tasks, such as document classification or clustering.

For instance, by stemming or lemmatizing words like "visiting", "visited", and "visits" to their base form "visit", we can treat these words as a single feature, rather than separate features. This reduces the dimensionality of the feature space and can lead to more effective analysis. Stemming and lemmatization are widely used techniques in text preprocessing for data mining and natural language processing tasks. They help in reducing the complexity of the text data while retaining the essential semantic information, enabling more efficient and accurate analysis.

c. **Part-of-speech tagging:** Assigning grammatical tags (e.g., noun, verb, adjective) to words in a text can help extract more meaningful features and improve the accuracy of text mining tasks [27].
Here's an example of part-of-speech tagging for the sentence "John Doe visited St. Mary's Hospital on 2023-04-15 with a fever.":

John/NNP Doe/NNP visited/VBD St./NNP Mary/NNP 's/POS Hospital/NNP on/IN 2023-04-15/CD with/IN a/DT fever/NN ./.
In this example, each word in the sentence is followed by a slash and its corresponding part-of-speech tag:
- NNP: Proper noun, singular
- VBD: Verb, past tense
- POS: Possessive ending
- IN: Preposition or subordinating conjunction
- CD: Cardinal number
- DT: Determiner
- NN: Noun, singular or mass
- .: Punctuation mark, sentence closer

By assigning these grammatical tags to the words in the sentence, we can extract more meaningful features for text mining tasks. For instance, knowing that "visited" is a verb (VBD) and "Hospital" is a proper noun (NNP) can help in named entity recognition tasks, such as identifying healthcare-related actions and entities. Part-of-speech tagging enables data mining algorithms to better understand the syntactic structure of the text, which can lead to more accurate analysis and insights. This is particularly useful in domains like healthcare, where understanding the relationships

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 7, July 2024

|DOI: 10.15680/IJIRSET.2024.1307090|

between entities (e.g., patients, healthcare providers, and medical conditions) is crucial for making informed decisions and improving patient outcomes.

d. Named entity recognition: Identifying and classifying named entities, such as persons, organizations, and locations, can provide valuable context and enable more targeted analysis [28].

Here's an example of named entity recognition for the sentence "John Doe visited St. Mary's Hospital on 2023-04-15 with a fever.":
[John Doe]_{Person} visited [St. Mary's Hospital]_{Organization} on [2023-04-15]_{Date} with a [fever]_{Symptom}.
In this example, the named entities in the sentence are enclosed in square brackets and labeled with their corresponding entity types as subscripts:

- Person: "John Doe"
- Organization: "St. Mary's Hospital"
- Date: "2023-04-15"
- Symptom: "fever"

By identifying and classifying these named entities, we can extract valuable contextual information from the text. This context can enable more targeted analysis and help in understanding the relationships between different entities.

For instance, knowing that "John Doe" is a person and "St. Mary's Hospital" is an organization can help in analyzing patient-hospital interactions. Similarly, identifying "2023-04-15" as a date and "fever" as a symptom can assist in tracking the temporal aspects of medical conditions and their associated symptoms. Named entity recognition is particularly useful in domains like healthcare, where understanding the relationships between patients, healthcare providers, medical conditions, and treatments is essential for making informed decisions and improving patient outcomes. By extracting and classifying named entities from unstructured text data, such as medical records or clinical notes, healthcare professionals can gain valuable insights and make data-driven decisions to enhance the quality of care.

Technique	Description
Bag-of-words	Represents text as a set of words, disregarding grammar and word order
TF-IDF	Reflects the importance of a word in a document relative to the entire corpus
Word embeddings	Maps words to dense vectors capturing semantic and syntactic relationships
Part-of-speech tagging	Assigns grammatical tags (e.g., noun, verb) to words in a text
Named entity recognition	Identifies and classifies named entities (e.g., persons, organizations) in text

Table 2: Common techniques used in feature engineering for text data [22-28]

C. Applying cosine similarity and feature engineering in data mining

By applying cosine similarity to the engineered feature vectors, data mining algorithms can identify contextually similar items across disparate data sources [15]. For example, in a customer support context, cosine similarity can be used to find similar support tickets or customer inquiries, enabling faster resolution and improved resource allocation.
The combination of cosine similarity and feature engineering enables data mining algorithms to match and group similar data points, even when they are expressed using different words or formats [16]. This is particularly useful when dealing with unstructured or semi-structured data from various sources, such as social media posts, customer reviews, or medical records.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 7, July 2024

|DOI: 10.15680/IJIRSET.2024.1307090|

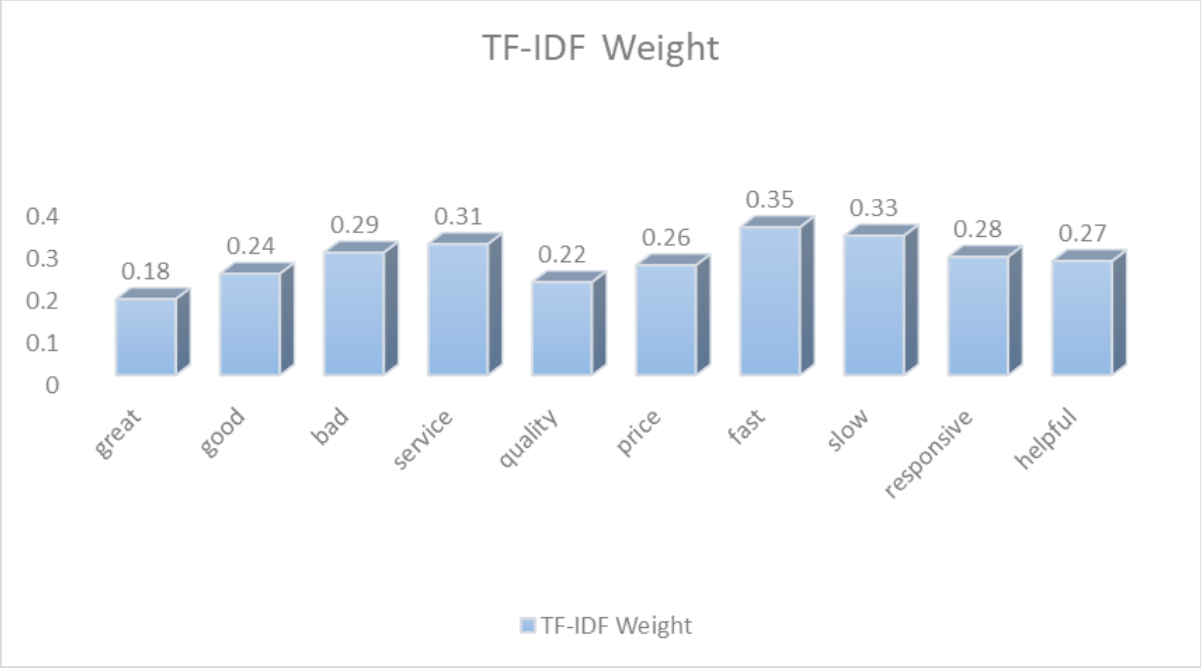

Figure 2: Sample dataset for TF-IDF feature importance in customer reviews [30]

In summary, cosine similarity and feature engineering are powerful tools in the data mining arsenal, enabling algorithms to effectively measure similarity and extract meaningful features from complex, high-dimensional datasets. By applying these techniques, businesses can uncover valuable insights and patterns that might otherwise remain hidden, ultimately leading to better decision-making and improved outcomes.

IV. BENEFITS OF ADVANCED DATA MINING TECHNIQUES

Advanced data mining techniques, such as context remembering, cosine similarity, and feature engineering, offer numerous benefits to businesses seeking to extract valuable insights from their data. These techniques enable organizations to effectively handle the challenges posed by the increasing volume, variety, and complexity of modern data sources.

A. Harnessing the full potential of data

By employing advanced data mining techniques, businesses can uncover hidden patterns, relationships, and trends within their data that might otherwise go unnoticed [17]. Context-aware data mining allows organizations to dive deeper into their data, considering the contextual relationships between data points and extracting insights that traditional data mining methods might miss. This ability to harness the full potential of data empowers businesses to make more informed decisions, optimize processes, and identify new opportunities for growth and innovation.

B. Deriving accurate and meaningful insights

By leveraging techniques like cosine similarity and feature engineering, context-aware data mining algorithms can identify similarities and group-related data items, even when they are expressed using different words or formats. Advanced data mining techniques provide more accurate and meaningful insights compared to traditional methods by considering the contextual relationships between data points [18]. This context-driven approach minimizes the impact of data inconsistencies and ensures that the derived insights are more reliable and relevant to the business's needs.

C. Dealing with complex and varied information

The proliferation of data sources, such as social media, and online platforms, has led to an exponential growth in the complexity and variety of information that businesses must process [19]. Advanced data mining techniques, including context remembering and feature engineering, enable organizations to effectively handle this complex and varied

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 7, July 2024

|DOI: 10.15680/IJIRSET.2024.1307090|

information by adapting to the unique characteristics of each data source in tandem with others. By considering the contextual nuances and leveraging techniques that can handle unstructured and semi-structured data, these advanced methods allow businesses to extract valuable insights from even the most challenging data sources.

Moreover, advanced data mining techniques can help businesses to:
1. Improve customer segmentation and targeting by identifying shared characteristics and preferences among customers, enabling personalized marketing and enhanced customer experiences.
2. Detect and prevent fraudulent activities by identifying unusual patterns and anomalies in transaction data, reducing financial losses and protecting brand reputation.
3. Optimize supply chain management by analysing data from various sources, such as inventory levels, shipping records, and supplier performance, to streamline operations and reduce costs.

As the importance of data-driven decision-making continues to grow, businesses that adopt advanced data mining techniques will be better positioned to capitalize on the opportunities presented by big data. By harnessing the full potential of their data, deriving accurate and meaningful insights, and effectively dealing with complex and varied information, organizations can gain a significant competitive advantage and drive long-term success.

V. CONCLUSION

In conclusion, data mining with context remembering, along with techniques like cosine similarity and feature engineering, has revolutionized the way businesses approach data analysis in the era of big data. By deriving reasoning from contextual search and resemblance, these advanced data mining techniques enable the unification and extraction of valuable insights from disparate and heterogeneous data sources. The ability to identify hidden patterns and relationships, even when data points are expressed using different words or formats, leads to more accurate and meaningful insights compared to traditional data mining methods. As businesses continue to generate and collect vast amounts of complex and varied information, the importance of context-aware data mining will only continue to grow. Embracing these advanced techniques will be crucial for organizations seeking to harness the full potential of their data and maintain a competitive edge in an increasingly data-driven world. Future research in this field should focus on further refining context-aware data mining algorithms, exploring new feature engineering techniques, and developing more efficient ways to handle the ever-growing volume and variety of data sources.

REFERENCES

[1] J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," Communications of the ACM, vol. 51, no. 1, pp. 107-113, Jan. 2008.
[2] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, 1993, pp. 207-216.
[3] J. Hirschberg and C. D. Manning, "Advances in natural language processing," Science, vol. 349, no. 6245, pp. 261-266, Jul. 2015.
[4] H. Chen, R. H. L. Chiang, and V. C. Storey, "Business intelligence and analytics: From big data to big impact," MIS Quarterly, vol. 36, no. 4, pp. 1165-1188, Dec. 2012.
[5] L. A. Kurgan and P. Musilek, "A survey of knowledge discovery and data mining process models," The Knowledge Engineering Review, vol. 21, no. 1, pp. 1-24, Mar. 2006.
[6] M. Ge, H. Bangui, and B. Buhnova, "Big data for internet of things: A survey," Future Generation Computer Systems, vol. 87, pp. 601-614, Oct. 2018.
[7] K. Niu, S. Huang, and L. Zhang, "Textual data mining based on topic modeling: A review," Expert Systems with Applications, vol. 168, p. 114379, Apr. 2021.
[8] G. Wang, Y. Jia, and Y. Sun, "A survey on natural language processing in data mining," in 2016 International Conference on Machine Learning and Cybernetics (ICMLC), 2016, pp. 1413-1418.
[9] L. Cao, "Data science: A comprehensive overview," ACM Computing Surveys, vol. 50, no. 3, pp. 1-42, Jun. 2017.
[10] A. Singhal, "Modern information retrieval: A brief overview," IEEE Data Engineering Bulletin, vol. 24, no. 4, pp. 35-43, Dec. 2001.
[11] A. Huang, "Similarity measures for text document clustering," in Proceedings of the Sixth New Zealand Computer Science Research Student Conference (NZCSRSC2008), 2008, pp. 49-56.
[12] A. Rajaraman and J. D. Ullman, "Mining of massive datasets," Cambridge University Press, 2011.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 7, July 2024

|DOI: 10.15680/IJIRSET.2024.1307090|

[13] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," arXiv preprint arXiv:1301.3781, 2013.
[14] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," Journal of Machine Learning Research, vol. 3, pp. 1157-1182, Mar. 2003.
[15] J. Han, M. Kamber, and J. Pei, "Data mining: Concepts and techniques," Morgan Kaufmann, 2011.
[16] S. Bird, E. Klein, and E. Loper, "Natural language processing with Python: Analyzing text with the natural language toolkit," O'Reilly Media, Inc., 2009.
[17] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," AI Magazine, vol. 17, no. 3, pp. 37-54, 1996.
[18] C. C. Aggarwal and C. K. Reddy, "Data clustering: Algorithms and applications," CRC Press, 2013.
[19] M. Chen, S. Mao, and Y. Liu, "Big data: A survey," Mobile Networks and Applications, vol. 19, no. 2, pp. 171-209, Apr. 2014.
[20] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," International Journal of Information Management, vol. 35, no. 2, pp. 137-144, Apr. 2015.
[21] A. Singhal, "Modern information retrieval: A brief overview," IEEE Data Engineering Bulletin, vol. 24, no. 4, pp. 35-43, Dec. 2001.
[22] J. Ramos, "Using TF-IDF to determine word relevance in document queries," in Proceedings of the first instructional conference on machine learning, 2003, vol. 242, pp. 133-142.
[23] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in Advances in neural information processing systems, 2013, pp. 3111-3119.
[24] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532-1543.
[25] C. Silva and B. Ribeiro, "The importance of stop word removal on recall values in text categorization," in Proceedings of the International Joint Conference on Neural Networks, 2003., 2003, vol. 3, pp. 1661-1666.
[26] M. F. Porter, "An algorithm for suffix stripping," Program, vol. 14, no. 3, pp. 130-137, 1980.
[27] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer, "Feature-rich part-of-speech tagging with a cyclic dependency network," in Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, 2003, pp. 252-259.
[28] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," Lingvisticae Investigationes, vol. 30, no. 1, pp. 3-26, 2007.
[29] J. Leskovec, A. Rajaraman, and J. D. Ullman, "Mining of massive datasets," Cambridge University Press, 2014.
[30] R. Feldman and J. Sanger, "The text mining handbook: Advanced approaches in analyzing unstructured data," Cambridge University Press, 2007.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details