



بِسْمِ اللّٰهِ  
الرَّحْمٰنِ الرَّحِیْمِ





# داوری مقالات

## وب کاوی

استاد : دکتر حمید طباطبایی

دانشجو : مجید فرهیخته اصل

آذر ۱۴۰۱



# Distributed Computing Frameworks for Supporting Big Data Analysis

IEEE.org | IEEE Xplore | IEEE SA | IEEE Spectrum | More Sites

SUBSCRIBE Cart Create Account Personal Sign In

IEEE Xplore<sup>®</sup> Browse ▾ My Settings ▾ Help ▾ Institutional Sign In

All [Search] ADVANCED SEARCH

Journal & Magazines > Big Data Mining and Analytics > Volume: 6 Issue: 2

## Survey of Distributed Computing Frameworks for Supporting Big Data Analysis

Publisher: TUP [Cite This] [PDF]

Xudong Sun, Yulin He, Dingming Wu, Joshua Zhexue Huang All Authors

83 Full Text Views

[R] [Share] [C] [Folder] [Bell]

Open Access

**Abstract**

**Document Sections**

- 1 Introduction
- 2 Distributed Technologies for Handling Big Data
- 3 Distributed Computing Frameworks for Big Data Analysis
- 4 Non-MapReduce Distributed Computing for Big Data Analysis
- 5 Evaluation

**Abstract:** Distributed computing frameworks are the fundamental component of distributed computing systems. They provide an essential way to support the efficient processing of big data on clusters or cloud. The size of big data increases at a pace that is faster than the increase in the big data processing capacity of clusters. Thus, distributed computing frameworks based on the MapReduce computing model are not adequate to support big data analysis tasks which often require running complex analytical algorithms on extremely big data sets in terabytes. In performing such tasks, these frameworks face three challenges: computational inefficiency due to high I/O and communication costs, non-scalability to big data due to memory limit, and limited analytical algorithms because many serial algorithms cannot be implemented in the MapReduce programming model. New distributed computing frameworks need to be developed to conquer these challenges. In this paper, we review MapReduce-type distributed computing frameworks that are currently used in handling big data and discuss their problems when conducting big data analysis. In addition, we present a non-MapReduce distributed computing framework that has the potential to overcome big data analysis challenges.

**Published in:** Big Data Mining and Analytics (Volume: 6, Issue: 2, June 2023)

**Page(s):** 154 - 169 **DOI:** 10.26599/BDMA.2022.9020014

**Date of Publication:** 26 January 2023 **Publisher:** TUP

**Electronic ISSN:** 2096-0654

**References:** Funding Agency:

Back to Results

### More Like This

**Efficient Distributed Database Clustering Algorithm for Big Data Processing**

2021 6th International Conference on Smart Grid and Electrical Automation (ICSGEA)

Published: 2021

**A Cost Minimization Model for a Multi-Component Product Closed Loop Supply Chain Considering Big Data Dimensions**

2021 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)

Published: 2021

Show More

BIG DATA MINING AND ANALYTICS  
ISSN 2096-0654 04/10 pp154-169  
Volume 6, Number 2, June 2023  
DOI: 10.26599/BDMA.2022.9020014

## Survey of Distributed Computing Frameworks for Supporting Big Data Analysis

Xudong Sun, Yulin He, Dingming Wu, and Joshua Zhexue Huang\*

**Abstract:** Distributed computing frameworks are the fundamental component of distributed computing systems. They provide an essential way to support the efficient processing of big data on clusters or cloud. The size of big data increases at a pace that is faster than the increase in the big data processing capacity of clusters. Thus, distributed computing frameworks based on the MapReduce computing model are not adequate to support big data analysis tasks which often require running complex analytical algorithms on extremely big data sets in terabytes. In performing such tasks, these frameworks face three challenges: computational inefficiency due to high I/O and communication costs, non-scalability to big data due to memory limit, and limited analytical algorithms because many serial algorithms cannot be implemented in the MapReduce programming model. New distributed computing frameworks need to be developed to conquer these challenges. In this paper, we review MapReduce-type distributed computing frameworks that are currently used in handling big data and discuss their problems when conducting big data analysis. In addition, we present a non-MapReduce distributed computing framework that has the potential to overcome big data analysis challenges.

**Key words:** distributed computing frameworks; big data analysis; approximate computing; MapReduce computing model

### 1 Introduction

In the era of big data, an overwhelming amount of data of various types is generated at all times from different channels, such as social networks, the Internet of Things, business transactions, finance networks, and personal media<sup>[1]</sup>. As a consequence, the explosive growth of global data has led to a fast increase in scales of accumulated data in data centers all over the world<sup>[2,3]</sup>. Through an analysis of big data, valuable

information and knowledge can be obtained, which benefits people from all walks of life and government and industry decision-makers<sup>[4,5]</sup>. In this scenario, distributed computing plays the most important role in storing, processing, and analyzing big data.

Distributed computing frameworks are the fundamental component of distributed computing systems. Using the divide-and-conquer strategy<sup>[6,7]</sup>, they provide an essential way to support the efficient processing of big data on clusters or cloud. In these frameworks, a big data file is partitioned into a number of small files called data block files, which are stored in a distributed fashion on the nodes of a cluster and managed by using a distributed file system such as Google File System (GFS)<sup>[8]</sup> and Hadoop Distributed File System (HDFS)<sup>[9-11]</sup>.

To process big data files on a cluster made of independent servers as nodes based on the shared-nothing architecture, local data block files are processed

\*Xudong Sun, Yulin He, Dingming Wu, and Joshua Zhexue Huang are with College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China. E-mail: sunxudong2016@email.szu.edu.cn; yulinhe@szu.edu.cn; dingming@szu.edu.cn; zx.huang@szu.edu.cn.

\*Yulin He and Joshua Zhexue Huang are also with Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), Shenzhen 518107, China.

\*To whom correspondence should be addressed.  
Manuscript received: 2022-06-15; accepted: 2022-06-28

# Distributed Computing Frameworks for Supporting Big Data Analysis



نتایج بررسی برای ژورنال با شناسه ۲۰۹۶۰۶۵۴

.Big Data Mining and Analytics

در تاریخ ۲۲ بهمن ۱۴۰۱

در فهرست سیاه وزارتین و دانشگاه آزاد اسلامی یافت نشد

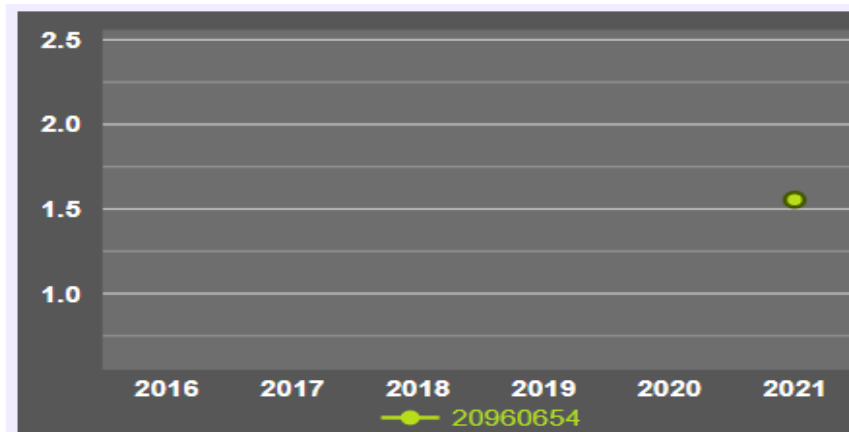
در فهرست مجلات نمایه JCR (دارای ضریب تاثیر) یافت نشد

در فهرست مجلات نمایه شده Master Journal List یافت شد

## BIG DATA MINING AND ANALYTICS

Address (Country) :	Coverage :
B605D, XUE YAN BUILDING, BEIJING, PEOPLES R.CHINA, 100084	Emerging Sources Citation Index

در فهرست مجلات نمایه شده Scopus یافت شد



## Big Data Mining and Analytics

ISSN	Publisher	SJR	H index	CiteFactor	Best Quartile
20960654	Institute of Electrical and Electronics Engineers Inc.	1.557	18	12.8	Q1

در فهرست مجلات نمایه شده DOAJ دسترسی آزاد یافت شد

Title	Publisher
Big Data Mining and Analytics	Tsinghua University Press

در فهرست مجلات نمایه شده PMC یا Medline یافت نشد

# Distributed Computing Frameworks for Supporting Big Data Analysis



Author Search Sources ?

## Sources

ISSN

ISSN: 2096-0654 x

**i** Improved Citescore x

We have updated the CiteScore methodology to ensure a more robust, stable and comprehensive metric which provides an indication of research impact, earlier. The updated methodology will be applied to the calculation of CiteScore, as well as retroactively for all previous CiteScore years (ie. 2018, 2017, 2016...). The previous CiteScore values have been removed and are no longer available.

[View CiteScore methodology.](#) >

### Filter refine list

[Clear filters](#)

### Display options

Display only Open Access journals

Counts for 4-year timeframe

No minimum selected

Minimum citations

1 result

[Download Scopus Source List](#) [Learn more about Scopus Source List](#)

All

View metrics for year: 2021

	Source title ↓	CiteScore ↓	Highest percentile ↓	Citations 2018-21 ↓	Documents 2018-21 ↓	% Cited ↓	
<input type="checkbox"/> 1	Big Data Mining and Analytics <i>Open Access</i>	12.8	96% 30/747	1,245	97	97	<a href="#">&gt;</a>

# Distributed Computing Frameworks for Supporting Big Data Analysis

توضیحات	ضعیف	متوسط	عالی	معیارها
کلمه کلیدی Big Data اضافه شود				آیا کلمات کلیدی مناسب هستند؟ کلمات کلیدی جدیدی پیشنهاد کنید.
عنوان مناسبی انتخاب شده بود				آیا عنوان راضی کننده است؟
چکیده جامع بود و به محتوای مقاله کاملا اشاره داشت				آیا چکیده، خلاصه جامعی از محتوای بیان شده ارائه میکند؟
همه موضوعات را پوشش داده بود و در مورد آنها توضیحات آورده شده بود				سرفصل ها به طور منطقی سازماندهی شده اند؟

# Distributed Computing Frameworks for Supporting Big Data Analysis

توضیحات	ضعیف	متوسط	عالی	معیارها
معرفی کمی طولانی بود				آیا اینتروداکشن شما را مجاب به خواندن بقیه متن کرد؟
				ایده‌ها و عناوین اصلی، به اندازه کافی مورد توجه قرار گرفتند؟
بخش‌ها و زیر بخش‌ها مختصر و مفید بود				آیا بخش‌ها و زیربخش‌های متفرقه، حجم متناسبی دارند؟
همه فرمت‌ها به درستی رعایت شده بود				آیا رفرنسها دارای فرمت صحیحی هستند و کل متن را در بر میگیرند؟ شامل نویسنده، عنوان، تاریخ، صفحه و آدرس اینترنتی هستند؟

# Distributed Computing Frameworks for Supporting Big Data Analysis

توضیحات	ضعیف	متوسط	عالی	معیارها
				آیا نویسنده در توضیحاتش از اصل (لوزی) پیروی کرده است؟
				اختصارات به درستی استفاده و لیست شده اند؟
تعداد تصاویر و جداول کم بود ولی توضیحات تصاویر آورده شده کامل بود				تصاویر و جداول (دارای لیبل گویا و ظاهر حرفه‌ای باشند، در متن به آنها ارجاع داده و در موردشان توضیح ارائه شده باشد).
اکثر پاراگراف‌ها طول مناسبی داشت و فقط تعداد کمی بیش از حد طولانی بود				آیا طول پاراگرافها صحیح است؟ (نه خیلی کوتاه و نه خیلی بلند)



# Distributed Computing Frameworks for Supporting Big Data Analysis

معیارها	عالی	متوسط	ضعیف	توضیحات
آیا نویسندگان به طور صحیح از زیرعنوانها استفاده کرده که بخشهای مختلف را مشخص نمایند؟	●			همه بخش ها به طور کامل آورده شده بود
آیا مطالب به گونه ای مرتب شده بود که منطقی، گویا و به راحتی قابل دنبال کردن باشد؟	●			بخش های سیر کاملی را طی می کرد
آیا قسمتی از متن وجود دارد که قابل حذف شدن باشد؟	●			در معرفی بعضی از بخش ها قابل حذف می باشد و اضافه بود
آیا خلاصه، به قسمت های کلیدی نتایج اشاره میکند؟	●			

# Distributed Computing Frameworks for Supporting Big Data Analysis

توضیحات	ضعیف	متوسط	عالی	معیارها
				آیا در متن، جداول و تصاویر، نقض حق تکثیر (کپی رایت) صورت گرفته؟
گرامر مشکل خاصی نداشت، فقط در بعضی کلمات به جای کلمه جمع از مفرد استفاده شده بود				تعداد خطاهای دستوری، نگارشی و نقطه گذاری (لطفا در متن علامت بزنید)
				آیا مطالعه جامع است؟
بله، در مورد پردازش داده های حجیم به روش های توزیع شده				آیا چیز جدیدی یاد گرفتید؟

# Distributed Computing Frameworks for Supporting Big Data Analysis

توضیحات	ضعیف	متوسط	عالی	معیارها
این مقاله در 26 January 2023 در IEEE به چاپ رسیده است <a href="https://ieeexplore.ieee.org/document/10026506/references#references">https://ieeexplore.ieee.org/document/10026506/references#references</a>				آیا کیفیت به اندازه ای بالا بود که قابلیت چاپ شدن در IEEE Communications Magazine را داشته باشد؟
۱۰				نمره ی کلی از ۱ تا ۱۰ (۱۰ یعنی بی نقص)
به صورت کامل تمامی مطالب را پوشش داده است				از چه چیز در این متن خوشتان آمد؟
تعداد تصاویر و جداول بیشتری استفاده نموده و برای توضیح بعضی از بخش ها از تصاویر استفاده کند روش های موجود را در نهایت در یک جدول جامع با هم مورد مقایسه قرار دهد بعضی از مطالب اضافی قابل صرف نظر کردن و حذف می باشد بعضی از پاراگراف ها بیش از حد طولانی شده اند				پیشنهادات شما برای ارتقاء متن

# Computational Intelligence in Web Mining



Innovative Trends in Computational Intelligence pp 197–215 | [Cite as](#)

## Computational Intelligence in Web Mining

[Dheeraj Kumar Singh](#), [Rohit Srivastava](#), [Tanupriya Choudhury](#) & [Anuj Kumar Yadav](#)

Chapter | [First Online: 30 November 2021](#)

258 Accesses

Part of the [EAI/Springer Innovations in Communication and Computing](#) book series (EAIISICC)

### Abstract

The increasing demand of web applications and social network websites generates a large volume of data for online accesses. Since web data stored across different web servers and online repositories have grown rapidly, understanding user's pattern and their content usage trends is essential for service providers. Web mining is an emerging technique in the field of computational intelligence. It is used to discover useful knowledge and insights from web data for a variety of applications such as target marketing, intrusion detection, web monitoring and recommendation, fake news analysis, etc. Web data contains heterogeneous data such as online documents, web structure data, web log, and user profile. Web content mining, web structure mining, and web usage mining are broad categories of web mining based on the type of data used in pattern extraction. This chapter describes basic functionalities of web mining and explores the state-of-the-art web mining techniques.

### Keywords

- Web mining
- Web data mining
- Web usage analysis
- Web recommendation
- E-content mining
- Web structure mining
- Computational intelligence in Web
- Web data analysis
- Web information retrieval
- web scraping

Access via your institution →

EUR 29.95  
Price includes VAT (Germany)

Chapter

- DOI: 10.1007/978-3-030-78284-9\_9
- Chapter length: 19 pages
- Instant PDF download
- Readable on all devices
- Own it forever
- Exclusive offer for individuals only
- Tax calculation will be finalised during checkout

Buy Chapter

▶ eBook	EUR 67.40
▶ Softcover Book	EUR 85.59
▶ Hardcover Book	EUR 121.98

[Learn about institutional subscriptions](#)

Sections    Figures    References

Abstract

[References](#)

[Author information](#)

[Editor information](#)

[Rights and permissions](#)

## Computational Intelligence in Web Mining

Dheeraj Kumar Singh, Rohit Srivastava, Tanupriya Choudhury & Anuj Kumar Yadav

Chapter

First Online: 30 November 2021

235 Accesses



Part of the [EAI/Springer Innovations in Communication and Computing](#) book series (EAIISICC)

[https://link.springer.com/chapter/10.1007/978-3-030-78284-9\\_9#Abs1](https://link.springer.com/chapter/10.1007/978-3-030-78284-9_9#Abs1)

### 1 Introduction

#### 1.1 General Overview of Web Mining

People, organizations, and the Web of Things (WoT) are generating incredible amount of data on the web every day. The size of data stored on the web server is often considered as a huge library. With the rapid development in web technology and Internet infrastructure, the World Wide Web (WWW) has become the most significant source of online data. The document structure of the online content is substantially more complex than offline documents. Computational intelligent approaches provide better solution for web mining system to enhance overall information retrieval performance. As in traditional data mining, the goal of web mining is to find and uncover valuable patterns from enormous web documents. Web data contains various types of data which include web archives information, hyperlink structure, web log file, and client profiles information [1]. All of the web data can be mined basically in three distinct categories, which are web content mining, web structure mining, and web usage mining.

# Computational Intelligence in Web Mining

توضیحات	ضعیف	متوسط	عالی	معیارها
کلمه کلیدی <b>Data Mining</b> را می توانست اضافه کند				آیا کلمات کلیدی مناسب هستند؟ کلمات کلیدی جدیدی پیشنهاد کنید.
عنوان مختصر ، کامل و گویا است و به داده کاوی اشاره می کند				آیا عنوان راضی کننده است؟
متاسفانه در خود مقاله بخش چکیده و کلمات کلیدی وجود ندارد اما در سایت <b>springer</b> این بخش ها وجود داشته و کامل می باشد				آیا چکیده، خلاصه جامعی از محتوای بیان شده ارائه میکند؟
سرفصل ها به ترتیب موضوعیت چیده شده و تمامی مطالب را پوشش می دهد				سرفصل ها به طور منطقی سازماندهی شده اند؟

# Computational Intelligence in Web Mining

توضیحات	ضعیف	متوسط	عالی	معیارها
معرفی جذابیت لازم جهت ادامه مطلب را داشت و به مسئله موجود و راه حل آن اشاره داشت				آیا اینترنت و اکشن شما را مجاب به خواندن بقیه متن کرد؟
عناوین مناسب برای هر بخش استفاده شده بود که به درستی به محتوای بخش ها اشاره می کرد				ایده ها و عناوین اصلی، به اندازه کافی مورد توجه قرار گرفتند؟
بعضی از بخش ها بیش از حد توضیح داده شد بود و می توانست به اشاره ای کوتاه به مطلب اکتفا کند				آیا بخشها و زیربخشهای متفرقه، حجم متناسبی دارند؟
منابع از فرمت استاندارد پیروی می کرد اما تعداد آنها کم بود				آیا رفرنسها دارای فرمت صحیحی هستند و کل متن را در بر میگیرند؟ شامل نویسنده، عنوان، تاریخ، صفحه و آدرس اینترنتی هستند؟

# Computational Intelligence in Web Mining

توضیحات	ضعیف	متوسط	عالی	معیارها
-				آیا نویسنده در توضیحاتش از اصل (لوزی) پیروی کرده است؟
تمامی اختصارات به درستی توضیح داده شده بود				اختصارات به درستی استفاده و لیست شده اند؟
تعداد تصاویر مناسب بود اما عناوین آنها بسیار مختصر بود و قابلیت توضیح بیشتر داشت - تعداد جداول نیز کم بود				تصاویر و جداول (دارای لیبل گویا و ظاهر حرفه‌ای باشند، در متن به آنها ارجاع داده و در موردشان توضیح ارائه شده باشد).
بعضی از پاراگراف‌ها بسیار طولانی بود و از حوصله خواننده خارج بود				آیا طول پاراگرافها صحیح است؟ (نه خیلی کوتاه و نه خیلی بلند)

# Computational Intelligence in Web Mining

توضیحات	ضعیف	متوسط	عالی	معیارها
بله عناوین به درستی انتخاب شده بود				آیا نویسندگان به طور صحیح از زیرعنوانها استفاده کرده که بخشهای مختلف را مشخص نمایند؟
بله مطالب از سیر مشخصی تبعیت می کرد و هدف آن مشخص بود				آیا مطالب به گونه ای مرتب شده بود که منطقی، گویا و به راحتی قابل دنبال کردن باشد؟
بعضی از بخش ها بیش از حد توضیح داده شده بود و قابلیت خلاصه شدن و یا حذف را دارا بود				آیا قسمتی از متن وجود دارد که قابل حذف شدن باشد؟
خلاصه به صورت کلی اشاره به وب کاوی کرده و تمامی کلمات کلیدی را پوشش نمیداد				آیا خلاصه، به قسمت های کلیدی نتایج اشاره میکند؟



# Computational Intelligence in Web Mining

توضیحات	ضعیف	متوسط	عالی	معیارها
شکل ها به صورت کلی بود که در وب یافت می شود و متعلق به شخص خاصی نیست لذا موضوعیتی ندارد				آیا در متن، جداول و تصاویر، نقض حق تکثیر (کپی رایت) صورت گرفته؟
مشکل خاصی در نگارش دیده نشد				تعداد خطاهای دستوری، نگارشی و نقطه گذاری (لطفا در متن علامت بزنید)
				آیا مطالعه جامع است؟
بله، با انواع روش های وب کاوی و مزایا و معایب و کاربرد آنها آشنا شدم				آیا چیز جدیدی یاد گرفتید؟

# Computational Intelligence in Web Mining

توضیحات	ضعیف	متوسط	عالی	معیارها
این بخشی از کتاب است اما قابلیت چاپ به صورت یک مقاله مروری را دارد <a href="https://link.springer.com/chapter/10.1007/978-3-030-78284-9_9#Abs1">https://link.springer.com/chapter/10.1007/978-3-030-78284-9_9#Abs1</a>				آیا کیفیت به اندازه ای بالا بود که قابلیت چاپ شدن در IEEE Communications Magazine را داشته باشد؟
۹				نمره ی کلی از ۱ تا ۱۰ (۱۰ یعنی بی نقص)
به صورت کامل فرآیند وب کاوی به روش های سنتی و جدید را توضیح داده است				از چه چیز در این متن خوشتان آمد؟
اضافه نمودن بخش های چکیده و کلمات کلیدی به مقاله اصلی همانند سایت از تعداد منابع بیشتر و جدیدتری استفاده کنند جدول مقایسه ای روش های وب کاوی را تهیه و تمامی آنها را در یک جدول مقایسه و مزایا و معایب هر کدام را به اختصار توضیح دهد از پاراگراف های کوتاهتری استفاده کنند				پیشنهادات شما برای ارتقاء متن

# World towards Advance Web Mining



Science and Education Publishing  
From Scientific Research to Knowledge

Keywords

Journal Home

For Authors

Online Submission

Current Issue

Archive

AJSS » Archive » Volume 3 » Issue 2 » Research Article

48392

IEWS

OPEN ACCESS

PEER-REVIEWED

## World towards Advance Web Mining: A Review

Shyam Nandan Kumar

M.Tech-Computer Science and Engineering, Lakshmi Narain College of Technology-Indore (RGPV, Bhopal), MP, India

Article

Metrics

Related Content

About the Authors

Comments

Follow the Authors

Abstract

1. Introduction

2. Web Mining

3. Data Mining vs. Web Mining

4. Web Usage Mining

5. Web Content Mining

6. Web Structure Mining

7. Semantic Web Mining

8. Web Mining Algorithms

9. Issues and Challenges in Web Mining

10. Web Mining Application

Abstract

With the advent of the World Wide Web and the emergence of e-commerce applications and social networks, organizations across the Web generate a large amount of data day-by-day. The abundant unstructured or semi-structured information on the Web leads a great challenge for both the users, who are seeking for effectively valuable information and for the business people, who needs to provide personalized service to the individual consumers, buried in the billions of web pages. To overcome these problems, data mining techniques must be applied on the Web. In this article, an attempt has been made to review the various web mining techniques to discover fruitful patterns from the Web, in detail. New concepts are also included in broad-sense for Optimal Web Mining. This paper also discusses the state of the art and survey on Web Mining that is used in knowledge discovery over the Web.

American Journal of Systems and Software, 2015, Vol. 3, No. 2, 44-61  
Available online at <http://pubs.sciepub.com/ajss/3/2/3>  
© Science and Education Publishing  
DOI:10.12691/ajss-3-2-3



SciEP  
Science & Education  
Publishing

## World towards Advance Web Mining: A Review

Shyam Nandan Kumar\*

M.Tech-Computer Science and Engineering, Lakshmi Narain College of Technology-Indore (RGPV, Bhopal), MP, India

\*Corresponding author: [shyamnandan.nec@gmail.com](mailto:shyamnandan.nec@gmail.com)

Received March 28, 2015; Revised April 05, 2015; Accepted April 16, 2015

**Abstract** With the advent of the World Wide Web and the emergence of e-commerce applications and social networks, organizations across the Web generate a large amount of data day-by-day. The abundant unstructured or semi-structured information on the Web leads a great challenge for both the users, who are seeking for effectively valuable information and for the business people, who needs to provide personalized service to the individual consumers, buried in the billions of web pages. To overcome these problems, data mining techniques must be applied on the Web. In this article, an attempt has been made to review the various web mining techniques to discover fruitful patterns from the Web, in detail. New concepts are also included in broad-sense for Optimal Web Mining. This paper also discusses the state of the art and survey on Web Mining that is used in knowledge discovery over the Web.

**Keywords:** data mining, www, web mining, cloud mining, web usage mining, web content mining, web structure mining, semantic web mining, web mining algorithm, knowledge discovery, information retrieval

**Cite This Article:** Shyam Nandan Kumar, "World towards Advance Web Mining: A Review." *American Journal of Systems and Software*, vol. 3, no. 2 (2015): 44-61. doi: 10.12691/ajss-3-2-3.

### 1. Introduction

Today, Web has turned to be the largest information source available in this planet. The Web is a huge, explosive, diverse, dynamic and mostly unstructured data repository, which supplies incredible amount of information, and also raises the complexity of how to deal with the information from the different perspectives of view – Users, Web service providers, Business analysts. The users want to have the effective search tools to find relevant information easily and precisely. To find the relevant information, users either browse or use the search service when they want to find specific information on the Web. When a user uses search service he or she usually inputs a simple keyword query and the query response in the list of pages ranked based on their similarity to the query. But due to the problems [1] with browser like: Low precision, which is due to the irrelevance of many of search results, and Low recall, which is due to the inability to index all the information available on the Web, users feel difficulty to find the relevant information on the web. The Web service providers want to find the way to predict the users' behaviors and personalize information to reduce the traffic load and design the Web site suited for the different group of users. The business analysts want to have tools to learn the users'/consumers' needs, like what the customer do and want. Mass customizing the information to the intended user or even to personalize it to individual customer is the big problem. Web mining is expecting tools or techniques to solve the above problems encountered on the Web. Sometimes, web mining techniques provide direct solution to above problems. On




the other hand, web mining techniques can be used as a part of bigger applications that addresses the above problems. Other related techniques from different research areas, such as database, information retrieval, and natural language processing, can also be used. Therefore, Web mining becomes a very hot and popular research field.

Web mining combines two of the activated research areas: *Data Mining* and *World Wide Web*. Data mining is the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data. It extracts the hidden predictive information from large database. With the widespread use of data and the explosive growth in their size, organizations are faced with the problem of information overload. The Web mining research relates to several research communities such as Database, Information Retrieval and Artificial Intelligence. Web mining, when looked upon data mining terms, can be said to have three operations of interests: *Clustering* (e.g. finding natural grouping of users, pages, etc.), *Association* (e.g. which URLs tend to be requested together), *Sequential Analysis* (e.g., the order in which URLs tends to be accessed). As in most real world problems, the clusters and associations in web mining do not have clear-cut boundaries and often overlap considerably. The unstructured feature of Web data triggers more complexity of Web mining. In the present time, it is not easy task to retrieve the desired information because of more and more pages have been indexed by search engines. So, this redundancy of resources has enhanced the need for developing automatic mining techniques on the WWW, thereby giving rise to the term "Web Data mining" [3]. Etzioni [4] came up with the question: Whether effective Web mining is feasible in practice? Today, with the tremendous growth of the data

# World towards Advance Web Mining

توضیحات	ضعیف	متوسط	عالی	معیارها
کلمات کلیدی تمامی مطالب مقاله را پوشش می دهد				آیا کلمات کلیدی مناسب هستند؟ کلمات کلیدی جدیدی پیشنهاد کنید.
عنوان کمی گمراه کننده است بهتر بود از عناوین زیر استفاده می نمود : Web Mining Research Study on Web Mining				آیا عنوان راضی کننده است؟
چکیده کامل بود و کلیات بحث را مشخص می کند				آیا چکیده، خلاصه جامعی از محتوای بیان شده ارائه میکند؟
بله سرفصل ها یک سیر تکاملی دارد و همه مطالب در این حوزه را پوشش می دهد				سرفصل ها به طور منطقی سازماندهی شده اند؟

# World towards Advance Web Mining

توضیحات	ضعیف	متوسط	عالی	معیارها
بخش معرفی نقشه راه مقاله را به درستی ترسیم می کند، مسئله و مشکل را مطرح کرده اما بیش از توضیح می دهد و از پاراگراف های طولانی و خسته کننده استفاده نموده است				آیا اینترنت و اکشن شما را مجاب به خواندن بقیه متن کرد؟
عناوین بخش ها به درستی انتخاب شده و کافی است				ایده ها و عناوین اصلی، به اندازه کافی مورد توجه قرار گرفتند؟
بخش ها و زیر بخش ها از حجم مناسبی برخوردارند				آیا بخشها و زیربخشهای متفرقه، حجم متناسبی دارند؟
فرمت صحیح منابع به درستی رعایت شده و بعضی از منابع به صورت رندوم تست شد و لینک ها معتبر می باشد				آیا رفرنسها دارای فرمت صحیحی هستند و کل متن را در بر میگیرند؟ شامل نویسنده، عنوان، تاریخ، صفحه و آدرس اینترنتی هستند؟

# World towards Advance Web Mining

توضیحات	ضعیف	متوسط	عالی	معیارها
-				آیا نویسنده در توضیحاتش از اصل (لوزی) پیروی کرده است؟
همه اختصارات به درستی و بلافاصله در خود متن مقاله آورده شده است				اختصارات به درستی استفاده و لیست شده اند؟
تعداد تصاویر و جداول خوب و توضیحات آنها کامل بود				تصاویر و جداول (دارای لیبل گویا و ظاهر حرفه‌ای باشند، در متن به آنها ارجاع داده و در موردشان توضیح ارائه شده باشد).
به جز در بخش معرفی بیشتر پاراگراف ها از اندازه مناسبی برخوردار بود				آیا طول پاراگرافها صحیح است؟ (نه خیلی کوتاه و نه خیلی بلند)

# World towards Advance Web Mining

توضیحات	ضعیف	متوسط	عالی	معیارها
عناوین بخش ها و زیر بخش ها به درستی انتخاب و گویای مطالب متن بخش مربوطه بود				آیا نویسنده به طور صحیح از زیرعنوانها استفاده کرده که بخشهای مختلف را مشخص نماید؟
بله سیر مقاله به درستی طی می شود				آیا مطالب به گونه ای مرتب شده بود که منطقی، گویا و به راحتی قابل دنبال کردن باشد؟
در بخش معرفی می توان بعضی از قسمت ها را حذف نمود و کلی تر مطالب را نوشت				آیا قسمتی از متن وجود دارد که قابل حذف شدن باشد؟
خلاصه به کلیات مقاله اشاره می کند				آیا خلاصه، به قسمت های کلیدی نتایج اشاره میکند؟

# World towards Advance Web Mining

توضیحات	ضعیف	متوسط	عالی	معیارها
-				آیا در متن، جداول و تصاویر، نقض حق تکثیر (کپی رایت) صورت گرفته؟
مقاله مشکلی از لحاظ نگارش و خطاهای دستوری نداشت				تعداد خطاهای دستوری، نگارشی و نقطه گذاری (لطفا در متن علامت بزنید)
بله، مقاله تمامی مطالب در حوزه داده کاوی و وب کاوی را پوشش می دهد				آیا مطالعه جامع است؟
بله با شاخه های مختلف کاوش Mining مانند خدمات کاوی، استخراج ابری، استخراج جریان کار و ... آشنا شدم				آیا چیز جدیدی یاد گرفتید؟



# World towards Advance Web Mining

توضیحات	ضعیف	متوسط	عالی	معیارها
با توجه به مشکلات بزرگ مقاله، تعداد منابع کم و قدیمی بودن مطالب ارائه شده قابلیت چاپ در مجله IEEE را ندارد				آیا کیفیت به اندازه ای بالا بود که قابلیت چاپ شدن در IEEE Communications Magazine را داشته باشد؟
۷				نمره ی کلی از ۱ تا ۱۰ (۱۰ یعنی بی نقص)
داده کاوی و وب کاوی را از حوزه های مختلف مورد بررسی قرار داده است				از چه چیز در این متن خوشتان آمد؟
از عنوان بهتری استفاده شود و از کلمات غلو آمیز پرهیز شود بخش معرفی را خلاصه تر و از پاراگراف های با اندازه مناسب استفاده شود از تعداد منابع بیشتری استفاده شود از توضیحات اضافه در بخش ها پرهیز شود				پیشنهادات شما برای ارتقاء متن



THANKS!

Majid Farhikhteh  
[majidfarhikhteh@gmail.com](mailto:majidfarhikhteh@gmail.com)