Appendix

Periodic progress reports (PPR)

Print Back

College : L. D. COLLEGE OF ENGINEERING, AHMEDABAD
StudentName : Akhawat Kalpit Ajeetkumar
EnrollmentNo : 150283116001 Department : Information Technology
MobileNo : 9727331128 Discipline : BE
Email : kalpitakhawat@gmail.com Semester : Semester 7

PPR Details

Periodic Progess Report : First PPR

Project : Pure Trending Content (Timeswen)

Status : Reviewed

1. What Progress you have made in the Project ?

We have done some analysis process on current system and some on UI and start to learn whatever technology is going to be used.

2. What challenge you have faced ?

How To Get The Data from all Sources?

3. What support you need ?

We required high computation power for parsing and removing duplication form crawled data.

4. Which literature you have referred ?

Some Data Analytics And Natural Language Processing research papers,blogs and journals

Comments

Comment by Internal Guide :
ok

Comment by External Guide :
None

Comment by HOD :
None

Comment by Principal :
None

Comment by University Admin :
None

Print Back

College : L. D. COLLEGE OF ENGINEERING, AHMEDABAD
StudentName : Bhatt Jigar Maneshbhai
EnrollmentNo : 150283116002 Department : Information Technology
MobileNo : 9638967123 Discipline : BE
Email : bhatt.jigar.214@ldce.ac.in Semester : Semester 7

PPR Details

Periodic Progess Report : Second PPR

Project : Pure Trending Content (Timeswen)

Status : Reviewed

1. What Progress you have made in the Project ?

We started researching about parallel processing architectures available for big data analysis. This research brought us on the cloud finally but it was very expensive so we had research on cloud implementation with local machines.

2. What challenge you have faced ?

How to implement cloud like architecture with local machines for achieving high computation power.

3. What support you need ?

Parallel Computation Power & a cloud setup. Hypervisor expert for cloud & cluster guidance.

4. Which literature you have referred ?

http://www.gtia.co.in/papers/e9bd6c2f-e271-44e6-8ef9-2fc36d79bbc31112012.pdf
http://www.communitysense.nl/papers/cacm06b.pdf
https://softwareengineering.stackexchange.com/questions/257661/patterns-for-creating-adaptive-web-crawler-throttling https://www.tensorflow.org/api_docs/ http://curator.apache.org/

Comments

Comment by Internal Guide :
None

Comment by External Guide :
None

Comment by HOD :
None

Comment by Principal :
None

Comment by University Admin :
None

Print Back

College : L. D. COLLEGE OF ENGINEERING, AHMEDABAD
StudentName : Fotariya Jimish Maheshbhai
EnrollmentNo : 150283116008 Department : Information Technology
MobileNo : 9712252863 Discipline : BE
Email : fotariyajimish@gmail.com Semester : Semester 7

PPR Details

Periodic Progess Report : Third PPR

Project : Pure Trending Content (Timeswen)

Status : Reviewed

1. What Progress you have made in the Project ?

We are learning about natural language processing & libraries like Natural language toolkit for finding how the machine can extract human language semantics. Were also checking the existing web crawlers & parser structure.

2. What challenge you have faced ?

How to Process data which have the same meaning but in different word structure.

3. What support you need ?

Expert for natural language processing.

4. Which literature you have referred ?

https://www.tensorflow.org/api_docs/
www.ijcscn.com/Documents/Volumes/vol4issue6/ijcscn2014040606.pdf
https://www.w3.org/standards/semanticweb/ semanticweb.org/ https://userpages.uni-koblenz.de/~staab/Research/Publications/sub-flairs2002.pdf https://docs.microsoft.com/en-us/azure/

Comments

Comment by Internal Guide :
None

Comment by External Guide :
None

Comment by HOD :
None

Comment by Principal :
None

Comment by University Admin :
None

Print Back

College : L. D. COLLEGE OF ENGINEERING, AHMEDABAD
StudentName : Solanki Hardik Dineshbhai
EnrollmentNo : 150283116026
Department : Information Technology
MobileNo : 8000334996
Discipline : BE
Email : hardik_solanki@live.com
Semester : Semester 7

PPR Details

Periodic Progess Report : Forth PPR

Project : Pure Trending Content (Timeswen)

Status : Reviewed

1. What Progress you have made in the Project ?

Weve started learning Algorithm Optimization techniques since Ranking of content is going to be done in a frequent way because of an uncertainty of timing in the world of trending news.

2. What challenge you have faced ?

We wanted to speed up our ranking algorithm but there is an issue of algorithm inconsistency. Good-trading content can lose rank because of speed up modification logics in the existing algorithm. So the final decision was taken was to design a new ranking algo. from scratch.

3. What support you need ?

Other experts asked before, plus a GraphQL & neo4j Expert.

4. Which literature you have referred ?

https://www.tensorflow.org/api_docs/ https://codesachin.wordpress.com/2015/11/14/k-means-clustering-with-tensorflow/ https://www.tensorflow.uk/tag/hierarchical-clustering/ http://mahout.apache.org/docs/0.13.1-SNAPSHOT/algorithms/linear-algebra/d-spca.html http://ieeexplore.ieee.org/abstract/document/58871/?reload=true www.nltk.org/

Comments

Comment by Internal Guide :

None

Comment by External Guide :

None

Comment by HOD :

None

Comment by Principal :

None

Comment by University Admin :

None

Patent search and analysis reports (PSAR)

GUJARAT TECHNOLOGICAL UNIVERSITY

(GTU)

INNOVATION COUNCIL (GIC)

Patent Search & Analysis Report

(PSAR)

Date of Submission : 31/08/2017

Dear **Akhawat Kalpit Ajeetkumar,**

Studied Patent Number for generation of PSAR	:	**17BE7_150283116001_1**

PART 1: PATENT SEARCH DATABASE USED

1. Patent Search Database used	:	Indian Patent Office database
Web link of database	:	http://ipindiaservices.gov.in/publicsearch/
2. Keywords Used for Search	:	Trending ,Content,news
3. Search String Used	:	pure trending content news algorithm
4. Number of Results/Hits getting	:	0

PART 2: BASIC DATA OF PATENTED INVENTION /BIBLIOGRAPHIC DATA

5. Category/ Field of Invention	:	
6. Invention is Related to/Class of Invention	:	Get Trending Data
6 (a) : IPC class of the studied patent	:	G06F17/30; G06F7/00; G06F17/27
7. Title of Invention	:	System and method for discovering story trends in real time from user generated content
8. Patent No.	:	
9. Application Number	:	13/856398
9 (a) : Web link of the studied patent	:	http://www.freepatentsonline.com/9235635.html
10. Date of Filing/Application (DD/MM/YYYY)	:	03/04/2013
11. Priority Date (DD/MM/YYYY)	:	
12. Publication/Journal Number	:	
13. Publication Date (DD/MM/YYYY)	:	
14. First Filled Country : Albania	:	

15. Also Published as

Sr.No	Country Where Filled	Application No./Patent No.
1		

16. Inventor/s Details.

Sr.No	Name of Inventor	Address/City/Country of Inventor
1	Ittiachen	Bangalore, IN

17. Applicant/Assignee Details.

Sr.No	Name of Applicant/Assignee	Address/City/Country of Applicant
1	Yahoo Inc	Sunnyvale, CA, US

18. Applicant for Patent is : Individual

PART 3: TECHNICAL PART OF PATENTED INVENTION

19. Limitation of Prior Technology / Art

Currently, the trends in these user generated streams are surfaced as “ebay”, “announce deal”, “sell skype”, because users write about the same topics differently. The current state of the art fails to cohesively analyze user-generated streams to account for the variance in terminology used across a diverse data set

20. Specific Problem Solved / Objective of Invention

The present invention provides a solution allowing a system to intelligently parse and identify key trending topics and store topics or stories for subsequent analysis and retrieval.

21. Brief about Invention

A method for identifying story trends includes identifying a set of words in a fixed size data stream based on a subword cache, and electronically determining at least one story trend associated with the set of words and electronically generating a story hash associated with the set of words. The method also includes storing the story hash in a story trend cache and updating the story trend cache according to the story hash, and retrieving one or more popular story topics according to the story trend cache. Machine readable media including program code that causes execution of a method for generating search results also are described.

22. Key learning Points

Trending Analysis

23. Summary of Invention

The present invention further comprises a system for discovering story trends. The system comprises a plurality of client devices and a plurality of data sources coupled to a network. The system further comprises a web server operable to receive and transmit data to and from the client devices and data sources. In one embodiment the web server may be further operable to receiving a request for stories from a user and provide a plurality of stories to the user.

24. Number of Claims : 13

25. Patent Status : Published Application

26. How much this invention is related with your IDP/UDP?

Not related to IDP/UDP, It's related to my area of interest

27. Do you have any idea to do anything around the said invention to improve it? (Give short note in not more than 500 words)

No. I don't have any idea about improvement on this. We have ideas of integrating this with other algorithms.

GUJARAT TECHNOLOGICAL UNIVERSITY

(GTU)

INNOVATION COUNCIL (GIC)

Patent Search & Analysis Report

(PSAR)

Date of Submission : 11/09/2017

Dear **Solanki Hardik Dineshbhai,**

Studied Patent Number for generation of PSAR	:	**17BE7_150283116026_2**

PART 1: PATENT SEARCH DATABASE USED

1. Patent Search Database used	:	Google Patents
Web link of database	:	https://patents.google.com/
2. Keywords Used for Search	:	Content Duplication detection,Content Duplication detection,Content Duplication detection
3. Search String Used	:	Content Duplication detection
4. Number of Results/Hits getting	:	7373

PART 2: BASIC DATA OF PATENTED INVENTION /BIBLIOGRAPHIC DATA

5. Category/ Field of Invention	:	
6. Invention is Related to/Class of Invention	:	1/1
6 (a) : IPC class of the studied patent	:	NA
7. Title of Invention	:	Method and system for detecting duplicate documents in web crawls
8. Patent No.	:	
9. Application Number	:	US09343511
9 (a) : Web link of the studied patent	:	https://patents.google.com/patent/US6547829B1/en?q=Content&q=Duplication&q=Detection
10. Date of Filing/Application (DD/MM/YYYY)	:	06/30/1999
11. Priority Date (DD/MM/YYYY)	:	
12. Publication/Journal Number	:	
13. Publication Date (DD/MM/YYYY)	:	
14. First Filled Country : Albania	:	

15. Also Published as

Sr.No	Country Where Filled	Application No./Patent No.
1		

16. Inventor/s Details.

Sr.No	Name of Inventor	Address/City/Country of Inventor
1	Dmitriy Meyerzon	NA
2	Srikanth Shoroff F	NA
3	Soner Terek	NA
4	Scott Norin	NA

17. Applicant/Assignee Details.

Sr.No	Name of Applicant/Assignee	Address/City/Country of Applicant
1	Microsoft Technology Licensing LLC	1288 Pear Ave, CA, US

18. Applicant for Patent is : Group

PART 3: TECHNICAL PART OF PATENTED INVENTION

19. Limitation of Prior Technology / Art

NA

20. Specific Problem Solved / Objective of Invention

Method and system for detecting duplicate documents in web crawls

21. Brief about Invention

A Web crawler application takes advantage of a document store's ability to provide a content identifier (CID) having a value that is a unique function of the physical storage location of a data object or document, such as a Web page. In operation, the crawler first tries to fetch the CID for a document. If the CID attribute is not supported by the document store, the crawler fetches the document, filters it to obtain a hash function, and commits the document to an index if the hash function is not present in a history table. If the CID is available from the document store, the CID is fetched from the document store. The crawler then determines whether the CID is present in the history table, which indicates whether an identical copy of the document in question has already been indexed under a different URL. If the CID is present, indicating that the document has already been indexed, the new URL is placed in the history file but the document itself is not retrieved from the document store, nor is it filtered again to obtain a CID. If the CID is not present in the history table, the full document is retrieved and indexed. The CID data structure is an extension of a known globally unique ID (GUID). Whereas the GUID is a 16-byte number, the CID comprises a 16-byte GUID plus an additional 6-byte number.

22. Key learning Points

Web crawling

23. Summary of Invention

The present invention provides an improved way to access documents (including Web pages, file system documents, e-mail messages, etc.) stored in one or more document stores on a computer network. For example, the invention could be used in a Web crawler application, mail server, directory service, or any system requiring indexing or one-way replication of a document store. The invention is particularly directed to a method and system for identifying duplicate documents in a document store, and using this information to avoid unnecessarily retrieving and processing such duplicates.

24. Number of Claims : 22

25. Patent Status : Other (Active)

26. How much this invention is related with your IDP/UDP?

Not related to IDP/UDP, It's related to my area of interest

27. Do you have any idea to do anything around the said invention to improve it? (Give short note in not more than 500 words)

No. I don't have any idea about improvement on this. We have ideas of integrating this with other algorithms.

GUJARAT TECHNOLOGICAL UNIVERSITY

(GTU)

INNOVATION COUNCIL (GIC)

Patent Search & Analysis Report

(PSAR)

Date of Submission : 14/09/2017

Dear **Bhatt Jigar Maneshbhai,**

Studied Patent Number for generation of PSAR	:	**17BE7_150283116002_3**

PART 1: PATENT SEARCH DATABASE USED

1. Patent Search Database used	:	Free Patents Online
Web link of database	:	http://www.freepatentsonline.com/
2. Keywords Used for Search	:	trend Ranking ,content Ranking,ranking
3. Search String Used	:	trend Ranking, content Ranking
4. Number of Results/Hits getting	:	1016

PART 2: BASIC DATA OF PATENTED INVENTION /BIBLIOGRAPHIC DATA

5. Category/ Field of Invention	:	
6. Invention is Related to/Class of Invention	:	1/1
6 (a) : IPC class of the studied patent	:	G06F17/30
7. Title of Invention	:	Content ranking based on user features in content
8. Patent No.	:	
9. Application Number	:	14/147789
9 (a) : Web link of the studied patent	:	http://www.freepatentsonline.com/9633119.html
10. Date of Filing/Application (DD/MM/YYYY)	:	01-06-2014
11. Priority Date (DD/MM/YYYY)	:	
12. Publication/Journal Number	:	
13. Publication Date (DD/MM/YYYY)	:	
14. First Filled Country : Albania	:	

15. Also Published as

Sr.No	Country Where Filled	Application No./Patent No.
1		

16. Inventor/s Details.

Sr.No	Name of Inventor	Address/City/Country of Inventor
1	Wexler	Santa Carla
2	Mike	USA

17. Applicant/Assignee Details.

Sr.No	Name of Applicant/Assignee	Address/City/Country of Applicant
1	Yahoo Inc	Sunnyvale, CA, US

18. Applicant for Patent is : Individual

PART 3: TECHNICAL PART OF PATENTED INVENTION

19. Limitation of Prior Technology / Art

NA

20. Specific Problem Solved / Objective of Invention

This pattern is about personalized news-reading experience . In order to provide a better news-reading experience

21. Brief about Invention

Methods, systems, and computer programs are presented for providing a personalized news stream to a user. One method includes an operation for identifying user features associated with a user. The user features include personal features and social features. The personal features are based on activities of the user and the profile of the user. The social features are based on information about social connections of the user. The method further includes operations for extracting content features from a corpus of content items, for identifying intersections between user features and content features, and for assigning weights to the content features from the corpus based on the identified intersections. A score for each content item is determined based on the content features and the respective weights of the content items. The content items are then ranked based on the scores. One or more of the ranked content items are displayed.

22. Key learning Points

Content ranking

23. Summary of Invention

The user features include personal features and social features, the personal features being based on activities of the user and based on a profile of the user, and the social features being based on information about social connections of the user. Further, the method includes operations for extracting content features from a plurality of content items, identifying intersections between user features and content features for the plurality of content items, and for assigning weights to the content features from the plurality of content items based on the identified intersections. Further, the method includes operations for determining scores for each content item based on the content features and respective weights of the content items, and for ranking the plurality of content items based on the scores. One or more of the ranked plurality of content items are then displayed for the user. In another embodiment, the operations of the method are executed by a processor.

24. Number of Claims : 20

25. Patent Status : Published Application

26. How much this invention is related with your IDP/UDP?

Not related to IDP/UDP, It's related to my area of interest

27. Do you have any idea to do anything around the said invention to improve it? (Give short note in not more than 500 words)

No. I don't have any idea about improvement on this. We have ideas of integrating this with other algorithms.

GUJARAT TECHNOLOGICAL UNIVERSITY

(GTU)

INNOVATION COUNCIL (GIC)

Patent Search & Analysis Report

(PSAR)

Date of Submission : 11/09/2017

Dear **Fotariya Jimish Maheshbhai,**

Studied Patent Number for generation of PSAR : **17BE7_150283116008_4**

PART 1: PATENT SEARCH DATABASE USED

1. Patent Search Database used	:	PatentScope (WIPO Patent Database)
Web link of database	:	http://patentscope.wipo.int/search/en/search.jsf
2. Keywords Used for Search	:	web crawling ,html parsing,content parsing
3. Search String Used	:	html parsing
4. Number of Results/Hits getting	:	488

PART 2: BASIC DATA OF PATENTED INVENTION /BIBLIOGRAPHIC DATA

5. Category/ Field of Invention	:	
6. Invention is Related to/Class of Invention	:	1/1
6 (a) : IPC class of the studied patent	:	G06F 17/30,G06N 99/00
7. Title of Invention	:	METHOD OF AND SYSTEM FOR CRAWLING A WEB RESOURCE
8. Patent No.	:	
9. Application Number	:	15326045
9 (a) : Web link of the studied patent	:	https://patentscope.wipo.int/search/en/detail.jsf?docId=US200949331&recNum=5&tab=Drawings&maxRec=488&office=&prevFilter=&sortOption=Pub+Date+Desc&queryString=FP%3A%28web+crawling%29
10. Date of Filing/Application (DD/MM/YYYY)	:	01/26/2015
11. Priority Date (DD/MM/YYYY)	:	
12. Publication/Journal Number	:	
13. Publication Date (DD/MM/YYYY)	:	
14. First Filled Country : Albania	:	

15. Also Published as

Sr.No	Country Where Filled	Application No./Patent No.
1		

16. Inventor/s Details.

Sr.No	Name of Inventor	Address/City/Country of Inventor
1	Damien Raymond JeanFranois	USA
2	Liudmila Alexandrovna	USA
3	Egor Aleksandrovich	USA
4	Pavel Viktorovich	USA
5	Ivan Semeonovich	USA
6	Arsenii Andreevich	USA
7	Gleb Gennadievich	USA

17. Applicant/Assignee Details.

Sr.No	Name of Applicant/Assignee	Address/City/Country of Applicant
1	YANDEX	EUROPE AG

18. Applicant for Patent is : Group

PART 3: TECHNICAL PART OF PATENTED INVENTION

19. Limitation of Prior Technology / Art

NA

20. Specific Problem Solved / Objective of Invention

method of setting up a crawling schedule

21. Brief about Invention

There is disclosed a method of setting up a crawling schedule, the method executable at a crawling server, the crawling server coupled to a communication network, the communication network having coupled thereto a first web resource server and a second web 5 resource server. The method comprises: appreciating a first new web page associated with the first web resource server; appreciating a second new web page associated with the second web resource server.

22. Key learning Points

Web crawling & its scheduling system

23. Summary of Invention

There is disclosed a method of setting up a crawling schedule, the method executable at a crawling server, the crawling server coupled to a communication network, the communication network having coupled thereto a first web resource server and a second web 5 resource server. The method comprises: appreciating a first new web page associated with the first web resource server; appreciating a second new web page associated with the second web resource server

24. Number of Claims : 22

25. Patent Status : Published Application

26. How much this invention is related with your IDP/UDP?

Not related to IDP/UDP, It's related to my area of interest

27. Do you have any idea to do anything around the said invention to improve it? (Give short note in not more than 500 words)

No. I don't have any idea about improvement on this. We have ideas of integrating this with other algorithms.

GUJARAT TECHNOLOGICAL UNIVERSITY

(GTU)

INNOVATION COUNCIL (GIC)

Patent Search & Analysis Report

(PSAR)

Date of Submission : 11/09/2017

Dear **Fotariya Jimish Maheshbhai,**

Studied Patent Number for generation of PSAR	:	**17BE7_150283116008_5**

PART 1: PATENT SEARCH DATABASE USED

1. Patent Search Database used	:	Google Patents
Web link of database	:	https://patents.google.com/
2. Keywords Used for Search	:	Clustering on news ,Clustering on news ,Clustering on news
3. Search String Used	:	Clustering on news
4. Number of Results/Hits getting	:	3668

PART 2: BASIC DATA OF PATENTED INVENTION /BIBLIOGRAPHIC DATA

5. Category/ Field of Invention	:	
6. Invention is Related to/Class of Invention	:	1/1
6 (a) : IPC class of the studied patent	:	NA
7. Title of Invention	:	Methods and apparatus for clustering news content
8. Patent No.	:	
9. Application Number	:	US10611269
9 (a) : Web link of the studied patent	:	https://patents.google.com/patent/US7568148B1/en?q=news&q=duplicate
10. Date of Filing/Application (DD/MM/YYYY)	:	06/30/2003
11. Priority Date (DD/MM/YYYY)	:	
12. Publication/Journal Number	:	
13. Publication Date (DD/MM/YYYY)	:	
14. First Filled Country : Albania	:	

15. Also Published as

Sr.No	Country Where Filled	Application No./Patent No.
1		

16. Inventor/s Details.

Sr.No	Name of Inventor	Address/City/Country of Inventor
1	Krishna Bharat	NA
2	Michael Curtiss	NA
3	Michael Schmitt	NA

17. Applicant/Assignee Details.

Sr.No	Name of Applicant/Assignee	Address/City/Country of Applicant
1	Google Inc	Mountain View, CA

18. Applicant for Patent is : Group

PART 3: TECHNICAL PART OF PATENTED INVENTION

19. Limitation of Prior Technology / Art

NA

20. Specific Problem Solved / Objective of Invention

clustering news content

21. Brief about Invention

Methods and apparatus are described for scoring documents in response, in part, to parameters related to the document, source, and/or cluster score. Methods and apparatus are also described for scoring a cluster in response, in part, to parameters related to documents within the cluster and/or sources corresponding to the documents within the cluster. In one embodiment, the invention may identify the source; detect a plurality of documents published by the source; analyze the plurality of documents with respect to at least one parameter; and determine a source score for the source in response, in part, to the parameter. In another embodiment, the invention may identify a topic; identify a plurality of clusters in response to the topic; analyze at least one parameter corresponding to each of the plurality of clusters; and calculate a cluster score for each of the plurality of clusters in response, in part, to the parameter.

22. Key learning Points

Items clustering

23. Summary of Invention

"Methods and apparatus are described for scoring documents in response, in part, to parameters related to the document, source, and/or cluster score. Methods and apparatus are also described for scoring a cluster in response, in part, to parameters related to documents within the cluster and/or sources corresponding to the documents within the cluster. In one embodiment, the invention may identify the source; detect a plurality of documents published by the source; analyze the plurality of documents with respect to at least one parameter; and determine a source score for the source in response, in part, to the parameter. In another embodiment, the invention may identify a topic; identify a plurality of clusters in response to the topic; analyze at least one parameter corresponding to each of the plurality of clusters; and calculate a cluster score for each of the plurality of clusters in response, in part, to the parameter.

Additional aspects of the present invention are directed to computer systems and to computer-readable media having features relating to the foregoing aspects."

24. Number of Claims : 28

25. Patent Status : Published Application

26. How much this invention is related with your IDP/UDP?

Not related to IDP/UDP, It's related to my area of interest

27. Do you have any idea to do anything around the said invention to improve it? (Give short note in not more than 500 words)

No. I don't have any idea about improvement on this. We have ideas of integrating this with other algorithms.

Conclusion and flow

So how all things will proceed?

Before fetching data, we will need to have the targets to get data from. For that we will be gone list out most popular news feeders around the world.

There will be three main phases of the flow

- Crawling content from Internet and load to the data center.
- Parsing and indexing content and compute relevancy.
- Publishing content on the portal server.

Finding information by crawling.

We use software known as “web crawlers” to discover publicly available web pages. The most well-known crawler is called “TIMESWENBOT.” Crawlers look at web pages and follow links on those pages, much like you would if you were browsing content on the web. They go from link to link and bring data about those web pages back to TIMESWEN’s servers.

The crawl process begins with a list of web addresses from past crawls and site-maps provided by website owners. As our crawlers visit these websites, they look for links for other pages to visit. The software pays special attention to news sites, changes to existing sites and dead links.

Computer programs determine which sites to crawl, how often, and how many pages to fetch from each site. TIMESWEN doesn't accept payment to crawl a site more frequently for our web search results. We care more about having the best possible results because in the long run that’s what best for users and, therefore, our business.

We crawl under restriction policies like robot.txt. Most websites don’t need to set up restrictions for crawling, indexing or serving, so their pages are eligible to appear in search results without having to do any extra work. That said, site owners have many choices about how TIMESWEN crawls and indexes their sites through Webmaster Tools and a file called “robots.txt”. With the robots.txt file, site owners can choose not to be crawled by bot, or they can provide more specific instructions about how to process pages on their sites.

Organizing information by indexing.

The web is like an ever-growing public library with billions of books and no central filing system. TIMESWEN essentially gathers the pages during the crawl process and then creates an index, so we know exactly how to look things up. Much like the index in the back of a book, the TIMESWEN index includes information about words and their locations. When you search, at the most basic level, our algorithms look up your search terms in the index to find the appropriate pages.

The search process gets much more complex from there. When you search for “dogs” you don’t want a page with the word “dogs” on it hundreds of times. You probably want pictures, videos or a list of breeds. TIMESWEN’s indexing systems note many different aspects of pages, such as when they were published, whether they contain pictures and videos, and much more. With the Knowledge Graph, we’re continuing to go beyond keyword matching to better understand the people, places and things you care about.

We focus that you always get the Relevant content, that’s thy parsing is very important phase. After content will get parsed, then and then we can compute the relevancy of content according the algorithm that works like Facebook's news feed algorithm known as Edge-Rank.