International Journal of Management, IT & Engineering
Vol. 7 Issue 3, March 2017,
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gage as well as in Cabell's Directories of Publishing Opportunities, U.S.A

SAP HANA SEARCHING TECHNIQUE – ONE OF THE FASTEST WAYS OF SEARCHING ENTERPRISE DATA

Shirish Joshi

Abstract

Curiosity is human being's basic characteristic. Due to this curiosity various innovations have happened till date and still are happening. One of the forms of this habit is 'to search'! Day in-day out we require so much information for some or the other thing for which we ask for help, era before computers we referred books and teachers and friends but now it's just Google. Searching on Google has become integral part of our life irrespective of the profession you are in.

Have you ever thought what exactly goes behind this 'search engine' thing? Simply Google and you'll get the answer. Today there are various search engines available in the market for internet surfing but what if your business landscape needs one. How will you go about it? First of all, why is the search engine required in your business landscape?

Let's assume your business is in retail industry with turnover about a crore per annum. Products manufactured by your company are used by lakhs people on daily basis only in India. Your marketing and sales team are doing great job of spreading the network every minute using all possible mediums. Not only on ground front but also on internet, your company has grabbed the attention of customers by using online portal and e-shopping sites. Being a successful business, you need reviews, feedback about each product from customers and end users. In earlier days you have used feedback forms from the dealers about the customer review and now you have a helpline which takes feedback from customer on time to time basis. Other than this, there is a huge social media space where you get feedback from customer, end-user whether you ask for it or not. This feedback is in the form of Face book posts, tweets, Instagram images, emails, blogs,

reviews on e-shopping portal and companies' online portal. Feedback given is in the form of text containing emoticons, images, videos, google map links, regional language and what not.

Now it's a big question in front of your IT team to collect this data in one system and format, cleanse and scrutinize itaccording to the positive, negative or neutral feedback. Once the data is sorted and organized,use it to make reports for the higher management to give exact details about how many customers are happy, satisfied, not happy, dissatisfied about your products, service. IT team's difficulty increases from the point of collection of such diversified data up to the storage and structuring of it. Social media data is in the forms of strings/text, emoticons, images, short texts, typos, videos, audios, pdfs, maps.Which database understands images and videos and pdfs? In which type of table you will store it? How will you categorize it with primary key and foreign key combination? How will you create mappings between your product master data and it's feedback against a geographical area? How will you store google map link. So many questions huh!!

Keywords : SAP HANA, Full text Search, Fuzzy Search, Text Mining, Index

1. What is Enterprise data ?

Enterprise data is centralized data that is shared by the users of an organization, generally across departments and/or geographic regions. As the enterprise data is very important for all the parties involved, enterprises spend time and resources on careful and effective data modeling, solutions, security and storage of the data. It is important for the organization to precisely define, easily integrate and effectively retrieve data for both internal applications and external communication.

2. Introduction : Welcome to the world of SAP HANA!!

It's the one stop solution for all your data related, business landscape functional, technical problems related, background processing and social media data related problems. Biggest difficulty faced by today's businesses is of gathering social media unstructured data and processing it, is solved by SAP HANA in a snap. What is so special that SAP HANA offers and no other products in market offer?SAP HANA's massive processing speed and mammoth memory to store any form of data, be it structured or unstructured data(images, videos, map links

etc.).The way you google about any term on the internet, similar option is given by SAP HANA to dive intoyour own business data.Search engine given by SAP HANA offers various methods to analyze and mining the data grabbed from social media. Simple tabular format storage aids to process data faster and makes IT teams' life easy to build required reports in minutes.Moreover, management user now doesn't have to rely on IT team to make dashboard. In-built applications give full freedom to the management to search about anything and everything of the business.

3. How exactly this miracle happens in system?

SAP HANA has given various options of text processing like

- Native full-text and fuzzy search
- In-database text analytics
- Graphical modeling of search models
- Info Access – HTML5 UI toolkit and API for JavaScript

Data in any form can be stored in table with columnar storage. Column storage of data uses the less memory space as it stores all the values of a column in consecutive memory cells. This makes the data search faster and gives the output in all possible minutest timeframe.

Text processing can be categorized in three options

➢ Search : Search can be further classified as full-text indexing and fuzzy search.

➢ Text analysis : Text analysis does the processing of document to find out required values and store them in separate system generated table.

➢ Text mining : Text mining uses a combination of search and text analysis for basic operation and further does the processing based on vector space model.

4. SAP HANA server/platform Architecture

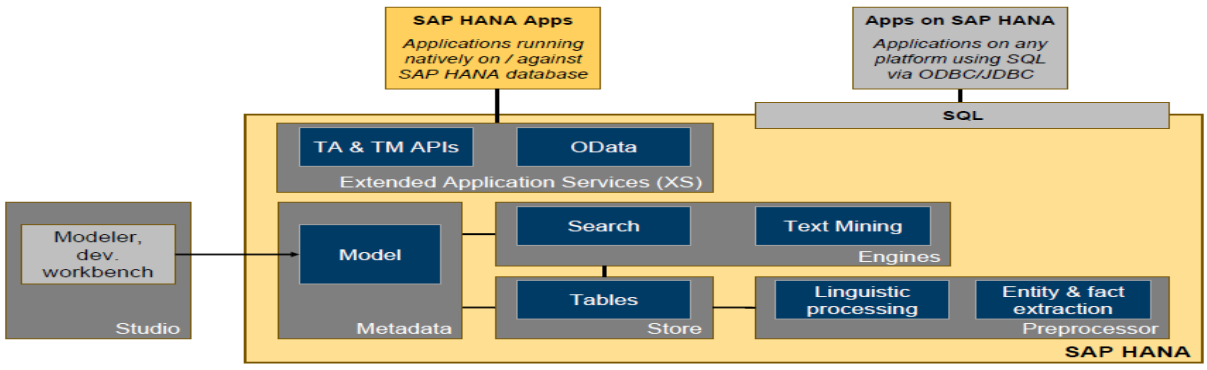


**image courtesy: scn.sap.com*

Above diagram is of SAP HANA server/platform. It is the main server which is responsible for all the processes running in landscape. This server takes care of all the transactions running in background and also, behaves as a database for various SAP and Non-SAP applications.

Preprocessor is heart of Text data processing. It is mainly responsible for pre-processing of unstructured data which comes from either various applications or SAP HANA models.It utilizes data which in stored in database tables or views.

Linguistic processing and Entity &Fact extraction play vital role in Text Analysis.
Gray box on the left side is SAP HANA Studio which is a tool given to do development, monitor, administrate and various functions of the SAP HANA server. This tool can be installed on desktop having OS like windows, Linuxor MacOS.

Metadata box, inside SAP HANA server, stores data about the various SAP HANA data models (calculation view, analytical views) created using SAP HANA studio. SAP HANA models are created to address various business requirements of transactional data generated every minute. Also, these models are used for creation of various reports and dashboards.

Engine's box contains Search & Text mining – these two engines take care of search term processing and text mining.

Extended Application Services (XS) engine is responsible for execution of HTML5 / UI5 based applications. These applications can be created using HTML5/UI5 languagewhich interact with core data stored in tables of SAP HANA database. These are lightweight applications specially designed for mobile, laptop and tablet platforms.

Search applications can be developed using HTML5/UI5 and data can be processed using built-in APIs (Application Programming Interface). These APIs can take care of TA (Text Analysis) & TM (Text Mining) of the structure, unstructured data.

Applications which are based on SQL can interact with SAP HANA server using JDBC/ODBC connection. Built-in APIs can take care of such third-party tool based applications and data processing can be done using SAP HANA power.

5. Search processing

Search can be further divided into two parts namely Full-text index search and Fuzzy search.

Full-text index method creates an index on the table which contains unstructured data in columnar form. This index is created automatically by the system. This index is non-accessible which means it is hidden for direct use but user can check the properties of table where you can get to know about it.Index creation, value updation in index, and deletion of index is handled by the system. Neither the developer nor the DBA (Database Administrator) has to do anything. If the table on which this index is created is dropped in future, then index is also dropped.

Values from source table stored in SAP HANA database are used for Index creation. If the user enters new search term in the application, index is not dropped but automatically updated.

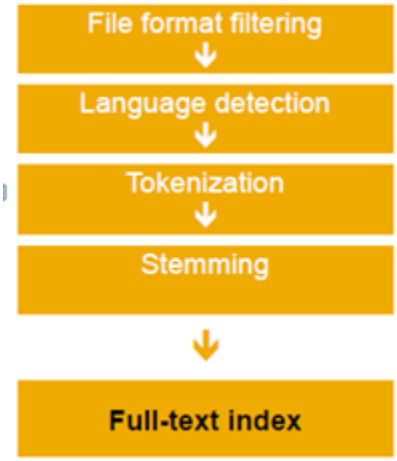
Usually, if the new value comes then the index is dropped and recreated on table. But in SAP system,index is not dropped but it gets updated with the delta functionality. Delta functionality is one of the core functionalities of SAP systems where only the newly created or changed record is added in the target space.

Search method works exactly like google. For example, if user enters ‘Tendulkar’ in google then result set shared by google contains all the possible links of Facebook, twitter, e-newspaper, LinkedIn, Instagram and so on where the word ‘Tendulkar’ is available. These links will give the data about not only Sachin Tendulkar name but all the persons whose name has word ‘Tendulkar’.

Basically, the search happens linguistic based and currently SAP supports 32 languages and many more are getting added. Wherever user enters a search term, sequence of characters or words doesn’t matter. System can still make out a valid word out of it and does the processing and returns maximum possible accurate results set.

But how exactly the system understands when to use the index created on source table. For using hidden column of the source table, command needs to be given to the system. When user enters any search term, the application in the background fires an SQL query which has WHERE clause with CONTAINS predicate. Only such SQL query can access the index column and gives the output faster.

Full-text indexing – How does it work step by step.



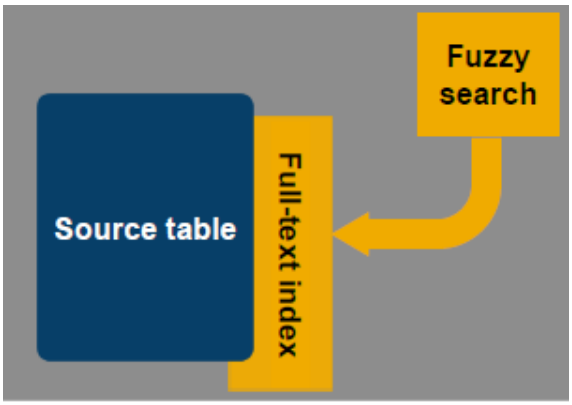
First of all file format of the unstructured data is identified and converted in text/HTML format. Further, it compares with the language pool defined in SAP HANA. Next step involves the defragmentation of sentence into different words in the form they are present in the sentence. Once words are separated, stemming (words are identified in the original form) is done and then index column is updated with new or changed values.

6. Other option available is Fuzzy search.

When you don't know the exact word to search for then fuzzy search helps you to find appropriate content. If you make typo error, it corrects it and searches the required content. At times it also suggests the right word while typing.

It uses 'Fuzzy math' i.e. statistics based search processing in the background. If at all you use a synonym for any word it gives the related content. For instance, if you want to search about USA sale of your products then, type just USA. It'll produce the results set which tells about USA, America and United States of America related sale of all the products.

If the system has built an index using Full-text index then, fuzzy search uses it else fuzzy math algorithms run and produce the output.

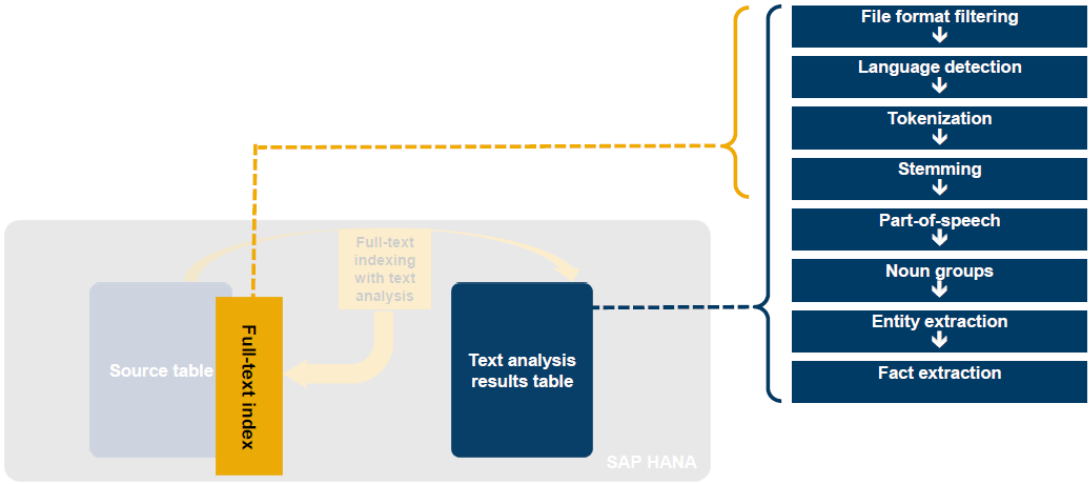


Above diagram explains the method a fuzzy search uses for output processing. As shown in the diagram, fuzzy search algorithm checks if index is available or not then continues the processing.

7. Text Analysis

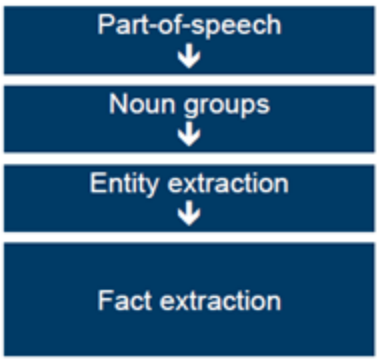
It is an option to the Full-text index method. Text analysis does the similar processing of unstructured data but an advantage is results are stored in separate table.

In case of Full-text index, index column is not accessible directly but, in case of Text Analysis, index column is stored as '$TA_Index Name'. This table is created by the system. All the values of index can now be used for further analysis in terms of dashboard reports and SAP HANA data models.



Above diagram shows the steps involved in Text Analysis processing.

From File format filtering up to Stemming, steps are same as of Full-text indexing.



Difference comes in the processing from 5th steps onwards.

Initial steps find out type of file by converting it into binary/html, identification of language, defragmenting the sentence and identifying the original word. Then

Part-of-speech: it finds out words as adjective, noun, singular or plural.

Noun groups: it identifies concept like text data or global piracy.

Entity extraction: it classifies pre-defined entity types.

e.g.: **Sachin Tendulkar** is a **cricketer**.

In this sentence, system will identify entity as 'Sachin Tendulkar – a person' and 'Cricket – organization@sports'.

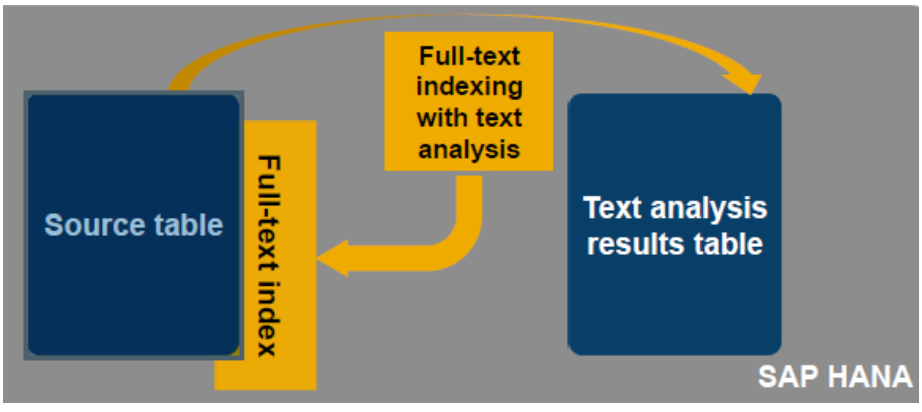
Fact extraction: it relates entities. This step classifies sentiments with topics.

e.g.: I **love** your **product**.

In this example, 'Love – strong positive sentiment' has occurred for the 'topic – Product'.

So, in this manner Text Analysis happens and final index/ output is prepared.

Ideally this analysis is used for social media data where end users have put in feedbacks on Facebook/twitter sites about your products.



So as shown in the above diagram, Text Analysis takes place and stores the result in $TA_<index name> table.

Basically it applies full linguistic and statistical techniques (i.e. natural language processing) to make sure the entities that get returned are correct.

Supported types for entity extraction in SAP HANA are:

- Who: People, job title, and national identification numbers
- What: Companies, organizations, financial indexes, and products
- When: Dates, days, holidays, months, years, times, and time periods
- Where: Addresses, cities, states, countries, facilities, Internet addresses, and phone numbers
- How much: Currencies and units of measure
- Generic concepts: text data, global piracy, and so on

Fact extraction is built on top of entity extraction and realized through rules that look for facts – relations between entities or states involving an entity.

Supported fact extraction

- **Voice of customer**
  - **Sentiments**: strong positive, weak positive, neutral, weak negative, strong negative, and problems
  - **Requests**: general and contact info
  - **Emoticons**: strong positive, weak positive, weak negative, strong negative
  - **Profanity**: ambiguous and unambiguous
- **Enterprise (only English language supported)**
  - Membership information, Management changes, Product releases, Mergers &acquisitions, Organizational information

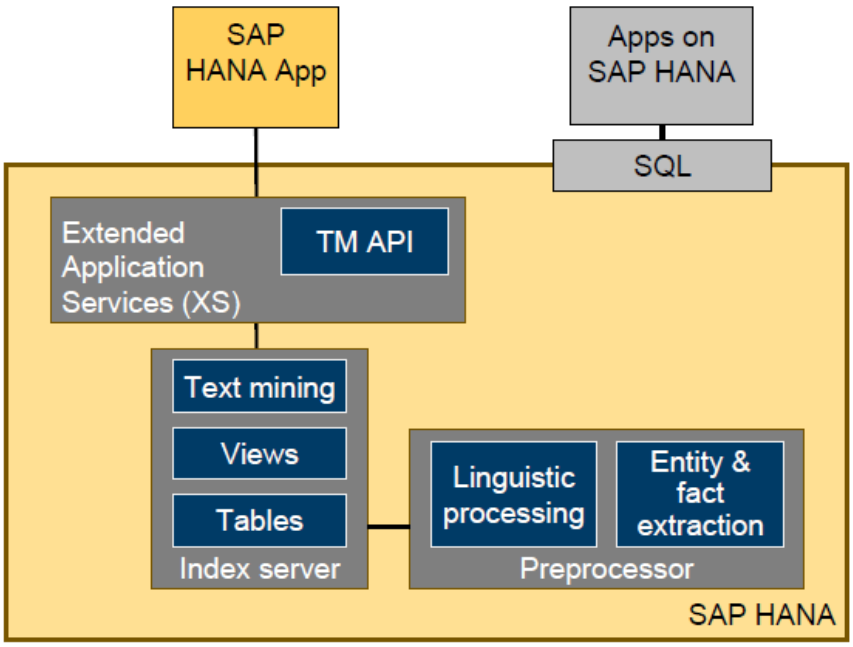
- **Public Sector (only English language supported)**
  - Action & travel events, Military units, Person alias, appearance, attribute relationships, spatial references, and Domain-specific entities.

Stored results of text analysis can further be used for following scenarios

Standard analytics	Create analytic view and calculation views on top E.g. Product 'X' searched purchased, returned on the company e-portal and other e-commerce portals.
Search-based applications	Create a search model and build a search UI with Info Access. Results can be used to navigate and filter search results. E.g. Search UI for Product master, Material master for Management users.
Data mining, predictive	Use R, SAP HANA Predictive Analysis Library (PAL) functions, graph, text mining... E.g. Clustering, time series analysis etc.

8. Text Mining

Text Mining is another option for Full-text indexing.It works at document level i.e. it compares two documents at the body level not line by line.It can be used as a post- process to text analysis. Processing happens based on a vector space model.



As shown in the above diagram, Text Mining takes place on the Index server. Native tables and views i.e. database tables / views which are stored in SAP HANA database are used while text

mining is taking place on data.SAP HANA provides a server-sideJavaScript or SQL interface for text mining.Text Mining can use all the approaches like full-text index and text analysis to do the data processing.Basics are done by text analysis and further mining can be done on that data.

9. Conclusion

Market is flooded with branded products like IBM's WebSphere, Informatica's Powercentre, Microsoft's Sharepoint offering enterprise search engine but not necessarily they fulfil the userrequirements. User wants the speed to get real-time data, accuracy in the unstructured data collected from internet, ease-to use the data to further process and make out meaningful dashboard to support the decision.

SAP HANA provides the processing speed (HANA database & server, columnar technology)– data accuracy(various methods to structure the unstructured data) – ease of use (can create various ad-hoc reports which can be consumed on any smart device). This makes the SAP HANA as one of the best products for searching the enterprise data.

10. References

- www.hana.sap.com
- www.scn.sap.com
- *SAP HANA : An Introduction by Be*rg Bjarne and Penny Silvia
- *Implementing SAP HANA by Don Loden*
- *SAP HANA Advanced Modeling by* Stefan Hartmann and Benedikt Engel
- *Implementing SAP HANA by* Jonathan Haun and Chris Hickman
- *SAP HANA Essentials 5th Edition by* Jeffrey Word
- SAP HANA Cookbook by Chandrasekhar Mankala and Ganesh Mahadevan V.