Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 11, November 2015, pg.227 – 232

RESEARCH ARTICLE

KNOWLEDGE MODELS FRAMEWORK USING TOPIC MAP

Prof. KumKum Bala

Computer Science and Engineering**,** Bharati Vidyapeeth's College Of Engineering, Lavale, Pune, India

itskumkum.bala@gmail.com

Abstract: For a large document corpus, the search engine is often the only way for newcomers to find what they are looking for. Traditionally, search has been driven by content keywords or by full-text indexing. The paper discuss the way of Integrating Knowledge Models built on top of Unstructured Text using frameworks like Topic Maps and Ontologies makes Organization understands the data well and can draw valuable Information easily.

I. Introduction

As organizations accumulate data from many different sources in the hopes of gaining competitive advantage with the big data, there is a risk associated with the process. Though newer analytics tools help organizations get insights from any data (something which was not available in the past), there will come a point where indiscriminate accumulation of data becomes a headache for these organizations. Yes, data storage is getting cheap and gives organizations an opportunity to accumulate data at a scale which they could never think about in the past. However, they also face the **following problems**

- Indiscriminate accumulation of data means higher costs for organizations.
- How to maintain the different unstructured data collected from different sources within the Organization or outside the Organization
- How to get the valuable Information from the data collected
- It is not always clear that the ROI for indiscriminate accumulation of data is always worth.
- Data governance will become a headache with such indiscriminate accumulation of data.
- It will not only give regulatory headache, it will also increase the cost of maintaining the data in a big way.

II. Solution

Integrating Knowledge Models built on top of Unstructured Text using frameworks like Topic Maps and Ontologies makes Organization understands the data well and can draw valuable Information easily.

The Knowledge Model Framework built using Topic Maps and Ontology, which will work on corpus of documents collected from either of the source and feed on to the system to generate KM Model. The KM framework will follow the plug and play architecture so that it can be easily integrated with different data source

Advantages of the solution are as follows:

- Get more attention from Search Engine.
- Plug and play architecture to integrate with different data sources
- Works on Unstructured data
- Models an Unstructured data to structured data
- Insights and clusters the data to draw meaningful information
- Attract unlimited new clients
- Data can be used for further analysis.

The Framework:

The framework based on Topic Maps with the help Ontology and domain model support to build Knowledge Models which will offer the alternative of indexing and searching against topic names, and then using topic occurrences to present links to all content related to the topics found by the search. The KM framework will follow the plug and play architecture so that it can be easily integrated with different data source.

Each topic in a topic map represents a single concept but can be assigned multiple names, allowing the topic map to store scientific and common-usage names, common misspellings, or translations for concept names. Ontology and Taxonomy support for validating the data and collecting the correct data from the source

Data collected from the data sources to be persisted in the Framework DB for processing. Web application to be created for showcasing the topic maps created.

III. Design

Architecture

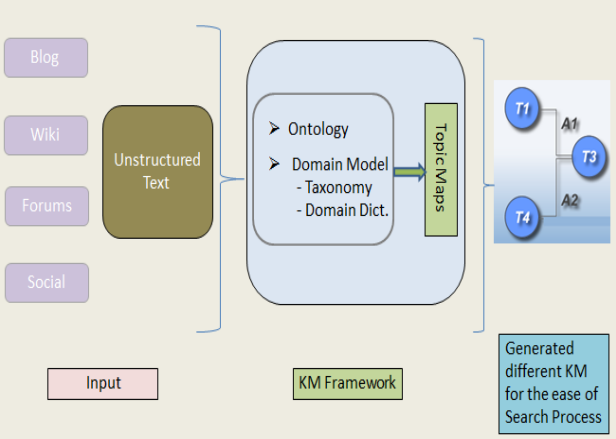

Figure 1: Solution Architecture

KM Framework consists of the different layers which will be responsible for Pre Processing of the data collected from the different sources and storing it into the framework DB.

Pre Processing Layer

The Preprocessing layer will work on the data collected to identify the accuracy and correctness of the data using Taxonomies containing different keywords

Topic Map Engine

The Framework will work on the data persisted in the DB and creates the topic maps. Topic Maps are a tool to organize information in a way that is optimized for navigation. It addresses the problem of info glut that we are facing. Too much information resolves eventually at no information, unless there are ways to filter and to extract efficiently the kind of information which is really needed. This problem has already been solved for printed material. Book indexes basically perform the same function, i.e. allowing readers to go directly to the portion of the document that is relevant to their information need. Topic Maps are the online equivalent of printed indexes, and it happens that they can do more: they are a powerful way to manage link information, such as glossaries, cross-references, thesauri, catalogs; they enable the merging of structured, unstructured information.

The fact that topic maps are now becoming an international standard is also an incentive for software vendors, who are now able to propose standardized tools to manage link databases

Taxonomy Layer

To validate the different topic maps created with the set of Keywords Organized to understand the different associations and classifications

Language			
Language	Java		
Language	Java	Keywords	
Language	Java	Keywords	for
Language	Java	Keywords	new
Language	Java	Keywords	switch
Language	Java	Keywords	assert
Language	Java	Keywords	default
Language	Java	Keywords	goto
Language	Java	Keywords	package
Language	Java	Keywords	synchronize
Language	Java	Keywords	const*
Language	Java	Keywords	float
Language	Java	Keywords	super
Language	Java	Keywords	while

Figure 2: Taxonomy

IV. Methodology

This framework is developed in JAVA. It follows the below methodology

- First part is to collect the data from the different Sources.
- Once data is collected each unstructured text is passed through Entity Extraction Module based on the GATE API to identify different Named Entity present in the document
- As soon as the different Named Entity is created out of the document it is validated against the domain layer using Taxonomy to identify the relevant Named Entities.
- Once the Entities are validated the Topic Map engine creates the Topic Maps out of it.

Framework consists of functions to interact with data collected from the different data sources and use the functions to create the Topic Map with the help of Taxonomy feed into the System.

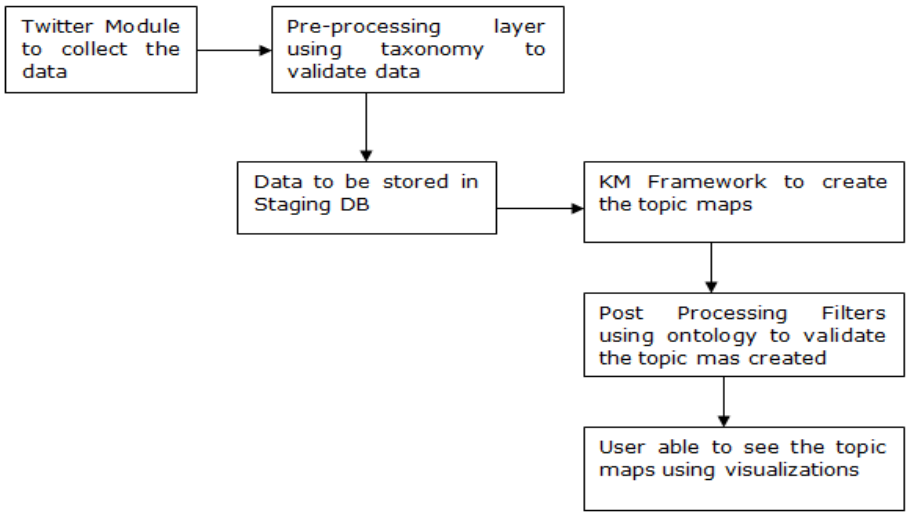

Figure 3: Taxonomy

Design Assumptions

- Users should have the well-defined Taxonomy in place.
- Users should have minimal automation testing knowledge.
- The architecture of this framework is arrived after analysis of the use cases to provide a time and cost saving solution for testing web applications.

Design Guidelines

The following are tips and guidelines followed

- Application should follow the plug and play Architecture to be used with any data source
- Application should be capable to display the Topic Maps created based on the User choice and Selection.
- Application can be run for different accounts.
- Application should not mix and Match data for different accounts

V. Technology Considerations

The following technology is used for the solution

Operating System	Windows
Languages	Java
Framework	TMAPI 1.0, Spring, J2EE, GATE
Database	MySQL
Web server	Tomcat 6.1.14
Other (Web Service)	Twitter 4J API

TMAPI

TMAPI is a programming interface for accessing and manipulating data held in a topic map. The TMAPI specification defines a set of core interfaces which must be implemented by a compliant application as well as a set of additional interfaces which may be implemented by a compliant application or which may be built upon the core interfaces.

TMAPI has been developed in an open process by developers working on topic map processors and topic map applications and placed into the public domain. There are no restrictions on its use.

Twitter4J API

With Twitter4J, you can easily integrate your - Java application with the Twitter service. (Twitter4J is an unofficial library)

We have chosen Twitter4J as it has following features:

- 100% Pure Java - works on any Java Platform version 1.4.2 or later
- Android platform and Google App Engine ready
- Zero dependency : No additional jars required
- Built-in Oath support
- Out-of-the-box grip support

GATE API

- GATE is open source software capable of solving almost any text processing problem
- A mature and extensive community of developers, users, educators, students and scientists
- A defined and repeatable process for creating robust and maintainable text processing workflows
- In active use for all sorts of language processing tasks and applications, including: voice of the customer; cancer research; drug research; decision support; recruitment; web mining; information extraction; semantic annotation
- The result of a multi-million R&D program running since 1995, funded by commercial users, the EC, BBSRC, EPSRC, AHRC, JISC, etc.
- used by corporations, SMEs, research labs and Universities worldwide
- The Eclipse of Natural Language Engineering, the Lucene of Information Extraction, the ISO 9001 of Text Mining
- A world-class team of language processing developers.

VI. Conclusion

The framework is used with Twitter Module to get the random 100 tweets based on the search term JAVA for which domain model is built. Different Named Entitles are extracted and validated and below output is generated using Topic Map Engine which shows different Topics, associations and Occurrences

Java Group				
Thread	Exception	OOPS	Keywords	Collection
Run	Try	Inheritance	do	list
Runnable	catch		while	array
			for	

Figure 3: Conclusion

VII. Future Extensibility

- This can enhanced to work with any data source connected
- The Topic Maps created can be further analyzed to know the Sentiments and different aspects of the Text Analytics
- Also this application can be extended to work on the structured data present in Organizations data bases
- The framework can be used to make the Knowledge Models within the organization from the unstructured data
- Ontology support can be implemented

References

[1] Diana Maynard, Valentin Tablan, Cristian Ursu, Hamish Cunningham and Yorick Wilks *Named Entity Recognition from Diverse Text Types* published in https://gate.ac.uk/sale/ranlp2001/maynard-etal.pdf

[2] Tin Huynh and Kiem Hoang *GATE framework based metadata extraction from scientific papers* published in Education and Management Technology (ICEMT), 2010 International Conference

[3] Garrido, A.L. IIS Dept., Univ. of Zaragoza, Zaragoza, Spain Buey, M.G. ; Escudero, S. ; Ilarri, S. ; Mena, E. ; Silveira, S.B. *TM-Gen: A Topic Map Generator from Text Documents* published in Tools with Artificial Intelligence (ICTAI), 2013 IEEE 25th International Conference.

[4] S. Pepper, "*The tao of topic maps*," in Proceedings of XML Europe, vol. 3, 2000.

[5] S. Pepper and G. Moore, "*XML Topic Maps (XTM) 1.0 - TopicMaps. org specification*," TopicMaps. Org Authoring Group, http://www.topicmaps.org/xtm, 2001

[6] Joshi, P., Chaudhary, S. , Kumar, V. *Information Extraction from Social Network for Agro-produce Marketing* published in Communication Systems and Network Technologies (CSNT), 2012 International Conference

[7] Andrei Mikheev, Marc Moens and Claire *Grover Named Entity recognition without gazetteers* published in EACL '99 Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics

