International Journal of Scientific Research & Engineering Trends
Volume 5, Issue 2, Mar-Apr-2019, ISSN (Online): 2395-566X

Extraction of Metadata and it's Parsing for PubMed articles

Aarti Mete Priyanka Kanthale Pritam Bhaye Rahul Subhedar Mrs. Shafali Gupta
Department of Computer Engineering
R.M.D Sinhgad College of Engineering
Pune , Maharashtra, India

Abstract- **Most of the data is in the form of articles and papers. Identifying the important articles form the millions of available papers is a need for today. Metadata extraction returns hidden information from the documents that are not in structured form to the one with semi - structured or standard form. This useful information contains authors, title, ID's, etc. about the article. Metadata is data (information) that is supposed to provide information from other data. This information is the basic information that can be further useful for the analysis of article.The extraction of metadata and the technique of parsing it have been discussed further.**

Keywords - Metadata, JSON, PMCID, Entrez.

I. INTRODUCTION

Metadata is a processed data that provides high level information about other data. The types of metadata are descriptive metadata, structural metadata, administrative metadata, referential metadata, statistical metadata and many more. Metadata is also known as the processed data that provides information about many aspects of data. It is used for summarization the basic information about data. This can make working with specific information easier.

JSON is the data type in python which is used to store the fields of metadata of the articles. JSON is a light mass processed data exchange format. This makes it easy for Analysis. It is also easy for machines to parse and generate useful information. It is a text format which is completely language independent. It also uses conventions.

Python provides with a library called Bio which comes with a package called Entrez. Entrez provides us with interface to connect the PubMed PMC records of metadata of the articles. Entrez comes with various functions for various uses of PMC articles. The esearch function gives us the records for the articles with their PMC IDs. These PMC IDs can then be passed to the summary function provided by Entrez.

The summary function of Entrez passes these PMC IDs to the PMC records to get the metadata JSON from them. These metadata JSONs contain various fields having the detailed information about the article. These JSONs can be then parsed to give the various independent records for the fields.

II. LITERATURE REVIEW

The author of [1] paper discusses techniques used in metadata mining. To teach computers how to analyze, understand and generate text, technologies are produced by natural language processing. There are various technologies. The types of situations where each technology is useful are the areas of interest. But the paper does not tell much about the techniques of text mining but only the survey.

The author of paper[2] show us that Metadata extraction is first step for computer. It helps to analyze unstructured text. It is done by identifying key phrases. It is also done sometimes by identifying relationships within text. To do this task process of pattern matching is used. It is useful to look for pre fixed series in text. Semantically they are interpreted and are then required pieces of information are entered into the database.

The author of paper[3] gives us the categorization of one or more category to free metadata. Categorization is supervised learning method because it is based on input output examples to classify new documents. Predefined classes are assigned to the text documents based on their content.

The author of paper[4] tells the representation of individual documents or groups of documents text flags are used to show metadata category and to show density used. Metadata mining puts large text and documental sources in visual levels. The other Papers are reference Papers for various ways and study of techniques that are used for the metadata extraction and its Parsing.

© 2019 IJSRET

III. METADATA EXTRACTION BASIC STAGES

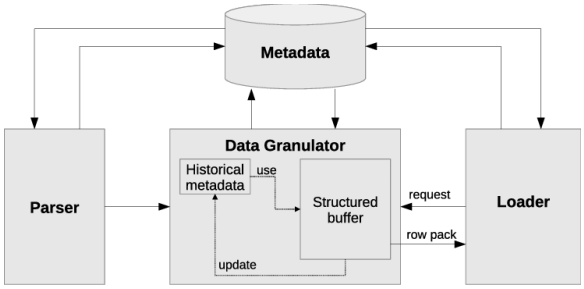
Usually the Metadata mining process has to undergo various steps. It is a very deep concept having numerous sub processing stages. Here are some main stages in Metadata mining. The steps in the Text mining are below:

Stage 1 Getting Records Using Entrez.esearch ():
This step is the foremost step in the text mining process. Here the PMC Records are taken into considerationthis step has to be carried out before carrying out any further step.

Stage 2 Passing the PMCID to Entrez.esummary ():
After the first step now the PMCID are passed on to the summary to get the records of metadata JSON for the particular article.

Stage 3: Parsing JSON:
The Metadata JSON that is received form the esummary is now passed on to the JSON parser where the various fields in the metadata JSON are now parsed to get the required fields

IV. PROPOSED FLOW



V. CONCLUSION

Thus, when working in python, the metadata for the PubMed articles can be covered using the various functions used in the Bio library of python where various APIs to get access to the PMC IDs and the metadata for the same can be carried out very easily.Also, these Metadata JSON when recovered from the esummary function can then be parsed using the JSONs and the various fields into that. The required fields can be stored to get the required metadata.

REFERENCES

[1] S. Bhatia and P. Mitra. Summarizing figures, tables, and algorithms in scientific publications to augment search results. ACM Transactions on Information Systems (TOIS), 30(1):3, 2012.

[2] R. Liu, W. Huang, and C. L. Tan.Extraction of vectorized graphical information from scientific chart images.In Document Analysis andRecognition, 2007.ICDAR 2007. Ninth International Conference on,volume 1, pages 521–525. IEEE, 2007.

[3] Y. Liu, K. Bai, P. Mitra, and C. L. Giles.Improving the table boundary detection in pdfs by fixing the sequence error of the sparselines. In Document Analysis and Recognition, 2009.ICDAR'09. 10thInternational Conference on, pages 1006–1010. IEEE, 2009.

[4] A. Sen, "Metadata management:past, present and future," Decision Sypport Systems, 2004, Vol.31, No.1, pp.151-173.

[5] S. Lawrence, C. L. Giles, and K. Bollacker , "Digital libraries andautonomous citation indexing," IEEE Computer, Vol.32, No.6, 1999,pp.67-71.

[6] H. Li, I. Councill, W. Lee, and C. L. Giles, "CiteSeerx: an architecture and web service design for an academic document search engine," Proceedings of the 15th international conference on World Wide Web, 2006, pp.883-884.

[7] W. Y. Arms. Digital Libraries.MIT Press, 2000.

[8] D. Bainbridge, K. J. Don, G. R. Buchanan, I. H. Witten,S. Jones, M. Jones, and M. I. Barr. Dynamic digital libraryconstruction and configuration. In Proc. European Conferenceon Digital Libraries, pages 1–16, 2004.

[9] D. Bainbridge, J. Thompson, and I.Witten. Assembling andenriching digital library collections. In Proc. Joint Conferenceon Digital Libraries, pages 323–334, 2003.

[10] D. Bainbridge, J. Thompson, and I. Witten, Assembling and enrichingdigital library collections, In Proc. Joint Conference on DigitalLibraries, pages 323–334, 2003.

[11] D. Bainbridge, K. J. Don, G. R. Buchanan, I. H. Witten, S. Jones, M. Jones, and M. I. Barr, Dynamic digital library construction andconfiguration, In Proc. European Conference on Digital Libraries, pages 1–16, 2004.

[12] The world fact book. http://www.cia.gov/cia/publications/factbook/, 2002.

[13] D. Tkaczyk and L. Bolikowski, "Workflow of Metadata Extraction from Retro-Born Digital Documents," in Towardsa Digital Mathematics Library, P. Sojka and T. Bouche, Eds.Masaryk University Press, 2011, pp. 39–44.

© 2019 IJSRET