IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 11, November 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.524

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 11, November 2024

|DOI: 10.15680/IJIRSET.2024.1311134|

Exploring Semantic Integration of Uncertain Web Data in a Multi-dialect Context for Kannada Documents

Kumarswamy H., Vinayaka V.M.

Department of CS&E, S.J.M. Institute of Technology, Visvesvaraya Technological University, Karnataka, India

Department of CS&E, Bangalore Institute of Technology, Visvesvaraya Technological University, Karnataka, India

ABSTRACT: This work proposes a model for building a Terminological Resource using web data tables and documents. This involves identifying symbolic concepts and quantities defined within the resource and represented in the columns of the table. The resource guides the annotation process by separating terminological and conceptual components, and by handling abbreviations and synonyms that might denote the same concept in a multi-dialect context. The built corpus consists of a generic part, representing the terminological structure for any existing documents or pages on web applicable to any domain, and a specific part for a particular field of interest. The work presents the model for resource building and its application in semantic annotation and querying of web documents.

KEYWORDS: Terminology, corpus, symbolic, annotation, synonyms, querying.

I. INTRODUCTION

A. Preamble

The internet is not just a collection of semi-structured documents linked together by hyperlinks; but it also houses a significant amount of technical and scientific documents, not only in one language; but in many dialects of a same language. Development of a software to retrieve such documents which are relevant using the semantic Web framework and language standards like XML and Resource Terminological Descriptions to enhance existing local data sources with terminological data extracted from Web documents require an AI model.

The system depends on a Terminological Resource, basically a database, which is divided into two parts: a generic set of concepts for data integration, and a specific set of concepts and terminology tailored to a particular application domain in different dialects of Kannada.

The software system comprises of two subsystems to build the Database:

1. Web Table Subsystem: This subsystem loads an XML/Resource Terminological Discerption data in JSON format at a cloud-based data warehouse with data tables extracted from Web documents, which are semantically annotated using Random Forest classifier.
2. Web Text Subsystem: This subsystem queries local data sources and the data warehouses using their API simultaneously employing CNN model to uniformly extract and detect keywords if the document is image pdf and SVM model for other documents to detect similarity of words in multiple dialects.

Thus, the Terminological Resource is built to provide approximate answers for user query using document resources in a consistent manner.

The Web Table subsystems involve four steps:

1. Retrieving and filtering relevant documents from the web for the application domain described in the annotation database by a human expert.
2. Semi-automatically extracting data tables from these documents.
3. Semantically annotating the extracted data tables using the Terminological Resource built using application of Random Forest on Kannada Dictionary, generating fuzzy annotations in a fuzzy Terminological Resource extension associated with the XML-represented data tables.
4. Building a JSON file to provide API for the application.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 11, November 2024

|DOI: 10.15680/IJIRSET.2024.1311134|

The Web Text Subsystem involves four steps:

1. Parsing the document which matches with the keyword using metadata of the document and preprocessing. It involves Data Preprocessing, after having imported the metadata and document title, removal of punctuations, stop words, etc. are done.
2. Keyword identification and Vectorizing the Keywords. Padding is done in order to determine a maximum length of the input to the Convolutional Neural Network.
3. Selection of a model, Word Embeddings, which is a representation of the data in a multidimensional space depending on similarity and dissimilarity between words, is created using Word2Vec.
4. The Word Embeddings are to be converted to a Matrix before fed as input to the CNN.

The Word Embeddings matrix output from the Embedding layer is the input to the Convolutional Layer, which performs the operation of Convolution.

Finally, the word which is similar is associated with the document source and a JSON file is built in the background.

The Web Table Subsystem is designed to precisely annotate target data tables from relevant web documents, requiring human intervention at each step to ensure accuracy. Whereas the Web Text Subsystem automatically annotates the documents and build a database in JSON format. This work primarily focuses on the second and third step: the vectorization of keywords and annotation method of the Web Text subsystem, which produces fuzzy database of Resource Terminological Descriptions and annotations. These annotations recognize imprecise numerical data and compute the semantic distance between table terms and text terms.

The flexible querying feature, allows querying these fuzzy annotations using JSON API. Its key features include retrieving both exact and semantically close answers and comparing selection criteria expressed as fuzzy sets with the fuzzy annotations.

II. LITERATURE SURVEY

[1] The paper speaks about advantage of diverse keyword extraction is that the coverage of the most topics of the voice communication fragment is maximized. The projected methodology for numerous keyword extraction returns in 3 steps, 1. Used to represent the distribution of the abstract topic for each word. 2. These topic models are used to determine weights for the abstract topics in each conversation fragment represented by βz. 3. The keyword list W = {w1, w2……wk}. Which covers a maximum number of the most important topics is selected by rewarding diversity, using an original algorithm introduced in this section.

[2] The paper presents the ONDINE system, designed to load and query a web-based data warehouse, guided by an Ontological and Terminological Resource (OTR). This data warehouse supplements local data sources with data tables extracted from web documents. The core process involves semiautomatic annotation of these tables using the OTR, resulting in an XML/RDF data warehouse with fuzzy RDF annotations. The paper also introduces a flexible querying system that allows simultaneous and uniform querying of both local sources and the data warehouse using SPARQL, retrieving approximate answers by comparing preferences expressed as fuzzy sets with the fuzzy RDF annotations.

[3] This paper presents an ontology-driven workflow that feeds and queries a data warehouse opened on the Web. Data are extracted from data tables in Web documents. As web documents are very heterogeneous in nature, a key issue in this workflow is the ability to assess the reliability of retrieved data. The main step of this method is to annotate and query Web data tables driven by domain ontology. This paper adopts a method to assess Web data table reliability from a set of criteria by the means of evidence theory. Finally, it shows how to extend the workflow to integrate, the reliability assessment step.

[4] In this paper the authors motivate why it is crucial to associate linguistic information with ontologies and why more expressive models, beyond the label systems implemented in RDF, OWL and SKOS, are needed to capture the relation between natural language constructs and ontological structures. They argue that in the light of tasks such as ontology-based information extraction (i.e., ontology population) from text, ontology learning from text, knowledge-based question answering and ontology verbalization, currently available models are not sufficient as they only allow us to associate literals as labels to ontology elements. Using literals as labels, however, does not allow us to capture additional linguistic structure or information which is definitely needed as they argue. In this paper they propose a

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 11, November 2024

|DOI: 10.15680/IJIRSET.2024.1311134|

model for linguistic grounding of ontologies called LexInfo. LexInfo allows us to associate linguistic information with respect to any level of linguistic description and expressivity to elements in ontology. LexInfo has been implemented as OWL ontology and is freely available together with an API. The authors have discussed the implementation of the LexInfo API, different tools that support the creation of LexInfo lexicons as well as some preliminary applications.

[5] There are a large number of ontologies currently available on the Semantic Web. However, in order to exploit them within natural language processing applications, more linguistic information than can be represented in current Semantic Web standards is required. Further, there are a large number of lexical resources available representing a wealth of linguistic information, but this data exists in various formats and is difficult to link to ontologies and other resources. The authors present a model that they call lemon (Lexicon Model for Ontologies) that supports the sharing of terminological and lexicon resources on the Semantic Web as well as their linking to the existing semantic representations provided by ontologies. Lemon can succinctly represent existing lexical resources and in combination with standard NLP tools we can easily generate new lexica for domain ontologies according to the lemon model. They have demonstrated that, by combining generated and existing lexica we can collaboratively develop rich lexical descriptions of ontology entities.

[6] Companies, governmental agencies and scientists produce a large amount of quantitative (research) data, consisting of measurements ranging from e.g. the surface temperatures of an ocean to the viscosity of a sample of mayonnaise. Such measurements are stored in tables for example spreadsheet samples and research reports. To integrate and reuse such data, it is necessary to have a semantic description of the data. However, the notation used is often ambiguous, making automatic interpretation and conversion to RDF or another suitable format difficult. For example, the table header cell “f (Hz)" refers to frequency measured in Hertz, but the symbol “f" can also refer to the unit farad or the quantities like force. Current annotation tools for this task mentioned in the paper either work on less ambiguous data or perform a more limited task.

[7] The authors present an algorithm, Nomen, for learning generalized names in text. Examples of these are names of diseases and infectious agents, such as bacteria and viruses. These names exhibit certain properties that make their identification more complex than that of regular proper names. Nomen uses a novel form of bootstrapping to grow sets of textual instances and of their contextual patterns. The algorithm makes use of competing evidence to boost the learning of several categories of names simultaneously.

[8] This paper claims to have improvised the results of ONDINE System[1] but our system uses the idea to retrieve documents having similar meaning in different dialects.

[9] The edit distance ed($w1$, $w2$) between two words $w1$ and $w2$ is the number of operations required to transform one of them into the other. The three primitive operations are 1) Substitution: changing one character to another in a word; 2) Deletion: deleting one character from a word; 3) Insertion: inserting a single character into a word. Given a keyword w, we let Sw,d denote the set of words $w\_$ satisfying ed($w,w\_$) $\leq d$ for a certain integer d.

III. TERMINOLOGICAL RESOURCE DATABASE BUILDING AND FUZZY ANNOTATIONS

A. The Conceptual Component of the Terminological Resource

The conceptual aspect of the Terminological Resource is its ontology. It comprises two main sections: a generic component, referred to as the core ontology, which holds the foundational concepts for data table semantic annotation, and a specific component, known as the domain ontology, containing concepts pertinent to a particular field. To grasp the structure of the core ontology, one must understand the semantic annotation process of data tables. A data table consists of columns made up of cells, which may need to be structured in a standardized format using tools like spreadsheets if not already formatted. The cells in a data table can contain terms or numerical values, often with measurement units. During semantic annotation, the cell contents are annotated to identify symbolic concepts or quantities represented in the columns, and to establish the semantic n-ary relationships between columns. The core ontology, therefore, includes three types of generic concepts: 1) simple concepts that encompass symbolic concepts and quantities, 2) unit concepts that include the units characterizing quantities, and 3) relations that facilitate the representation of n-ary relationships between simple concepts.[2]

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 11, November 2024

|DOI: 10.15680/IJIRSET.2024.1311134|

B. The Terminological Component of the Web base Text Resources

The terminological component of the Text Resource encompasses the terminology used in text documents relevant to the domain of interest. Each term, defined as a sequence of words in a specific language, has a label. Terms are categorized based on their source language dialect. A term represents a concept, with each term denoting at least one concept and potentially multiple concepts. The core terminology object property allows a term to refer to a concept, while the annotation functional data properties "Label" and "Language Dialect" associate a term with its label and language dialect, represented as a string.

This Terminological Resource is central to the system, which enables local data sources with annotated Web data tables and Text information. The semantic annotation of a data table involves two steps:

1. Identifying which Resource-defined relations are present in the data table.
2. Instantiating these relations by linking each row of the data table with a set of fuzzy annotation graphs.

Using text mining to build Terminological Resource of Text based resources involves parsing, tokenizing, and annotation of the keywords with language dialect *data taken from Kannada Dictionary.*

C. The Fuzzy Sets

A fuzzy set extends the concept of classical subsets by allowing elements to partially belong to the set, with membership degrees ranging from 0 (not part of the set) to 1 (fully part of the set). In classical subsets, elements either belong to a subset or its complement based on certain properties. Fuzzy sets can be continuous (CFS) if defined on a continuous domain including text documents, or discrete (DFS) if defined on a discrete domain including data tables. We use fuzzy set of annotations in order to avoid missing any relevant document being retrieved. By the query expansion through local context analysis and user relevance feedback the system is going to retrieve all the documents which match keywords of different dialects.

IV. PROPOSED SYSTEM

A. The Fuzzy Querying Method

The JSON querying system enables uniform querying of both local data sources and an XML/Terminological Resource data warehouse containing semantically annotated data tables from web documents. It utilizes the Ontological and Terminological Resource (OTR) for indexing and annotating data tables and allows users to express preferences in their queries and retrieve data close to their criteria by assessing the semantic proximity of the data using the OTR. Users interact through a single graphical user interface (GUI), which translates queries into SQL for relational sources and JSON API calls for the XML/Terminological Resource data warehouse. The final results are a combination of data from both sources, ordered by relevance. This subsystem also supports querying fuzzy annotations in document sources database.

B. Query to retrieve documents from original source

A JSON file maintains all the possible synonyms in different dialects mapping to the original resource. Query operates within a view corresponding to a specific relation in the Terminological Resource, characterized by its query-able attributes and definition. Each query-able attribute aligns with a simple concept in the view. This setup simplifies querying across diverse data sources for the end user. Users instantiate a view by selecting query-able attributes and their values. Queries can use continuous or discrete fuzzy sets to express preferences, retrieving exact and semantically close answers. When querying the XML/RDF data warehouse, the system handles fuzzy values by considering certainty scores and comparing fuzzy sets. The end user may set a threshold for certainty scores. Queries are "defuzzified" into JSON API calls, generating queries from the annotation selections.

C. The Construction of multiple internal queries to improvise document suggestions

To improve the retrieval results, multiple implicit queries are to be formulated for each word in the query fragment, using annotation of synonym in different dialect, with the keywords of each cluster from the previous annotations using CNN, as the search engine used in our system is not sensitive to word order in queries. Retrieval results must meet three essential criteria:

1. It must achieve at least the minimal acceptable certainty score defined by the querying person.
2. It has to satisfy all specified selection criteria including all the dialects.
3. It must assign a constant ID to each projected document results.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 11, November 2024

|DOI: 10.15680/IJIRSET.2024.1311134|

The process of suggesting document source using Terminological Resource built in cloud-based data warehouse involves three main steps:

1. Generating and Executing the Query: The system generates a query based on the query attributes of annotations built using different dialects and executes it within the data warehouse.
2. Extracting new keywords for Selected Attributes of a file and updating the Terminological Resource: When the values for the selected attributes are closer to the query keywords each fuzzy result results in Random Forest graph to determine how well the graph meets the selection criteria. Thus, adding the new source in to Terminological Resource and annotating it with different dialect keywords taken from Kannada Dictionary.
3. Projecting the documents with unique IDs based on the keyword attributes: The CNN and Random Forest classifier present in two of the subsystems run simultaneously for the projecting unique document sources in including those without data tables and those with data tables to provide the relevant data to the end user.

Keyword extraction from an document graph is performed through queries defined for each selection and projection attribute of the query. To measure how well a selection criterion is satisfied, two semantic aspects of fuzzy values—imprecision and similarity—are considered:

1. Imprecision: Classical measures are used to compare a fuzzy set representing preferences to a fuzzy set with a semantic of imprecision. These measures include:
   - Possibility Degree of Matching: Denoted by a specific measure, it indicates the potential alignment between the preferences and the fuzzy set.
   - Necessity Degree of Matching (N): This measure indicates how necessarily the preferences align with the fuzzy set.

2. Similarity: An adequation degree is used to compare a fuzzy set representing preferences to a fuzzy set with a semantic of similarity. This degree evaluates how similar the preferences are to the fuzzy set.

The end user may specify a threshold for the minimum acceptable certainty score, ensuring that only data meeting this level of certainty are retrieved. Since standard JSON APIs does not support fuzzy sets, the query is "defuzzified" before translation into JSON API calls. This approach allows any query to be transformed in to multiple JSON API calls before being used by the querying engine.

The query is generated automatically based on:
1. The synonym in different dialect of the relation represented by the view associated with the query.
2. The sets of projection and selection attributes specified in the query.

This elaborate process ensures that system provides accurate and relevant data to users, tailored to their specific preferences and criteria.

D. Web Text Subsystem

Web Text Subsystem is working to build a Terminological Resource data warehouse focused on integrating diverse documents of different dialects of Kannada retrieved while querying web documents. The emphasis on web tables is due to their frequent use for summarizing experimental data and their easier integration compared to text or plots. But since this part is taken care by Web Table Subsystem, the Web Text Subsystem skips any tables that are present the document. The central element in data integration in Web is the domain ontology, which details concepts, their relationships, and relevant terminology for a specific application domain. CNN extracts the keywords and the Web Text Subsystem annotates the keywords with dialect specific additional keywords and stores in the data ware house.

- Domain Ontology: Essential for describing concepts, relationships, and terminology for any given application domain. Once defined, it allows for the integration of new data into the warehouse seamlessly.
- Workflow: Involves steps to retrieve relevant documents using domain-specific keywords, semi-automatic extraction, and translation of data tables into a generic XML format, followed by semantic annotation of these tables based on the domain ontology.
- Core Ontology: Contains foundational concepts necessary for web table integration, including symbolic and numeric concepts and their relationships.
- Domain Ontology: Specific to the considered domain, containing domain-specific concepts.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 11, November 2024

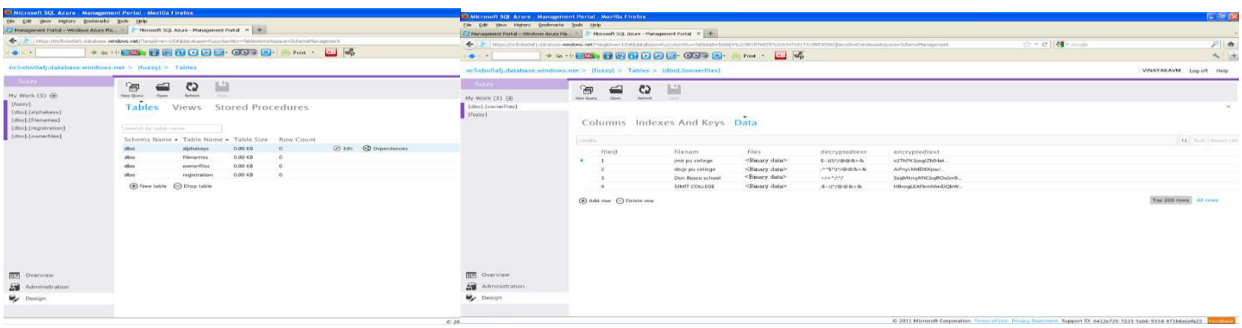
|DOI: 10.15680/IJIRSET.2024.1311134|

Workflow Steps:
1. Retrieve relevant web documents using keywords from the domain ontology and annotation of keyword dialects.
2. Extract additional keywords form the document retrieved and translate them into a generic XML format.
3. Semantically annotate these keywords by identifying semantic relations from the domain ontology in each dialect, generating Terminological Resource descriptions in JSON.
This approach allows for a uniform and structured way to interrogate and query the data warehouse across various application domains.

E. Query Execution Steps
1. Query Generation:
- The system generates a query based on the specified selection and projection attributes. This query is designed to interact with the XML/Terminological Resource data warehouse, leveraging the JSON API calls to locate relevant data.
2. Query Execution:
- The generated query is executed within the data warehouse. The warehouse uses the ontology to interpret the query and fetch the appropriate data.
3. Document Retrieval and additional keyword extraction:
- After executing the query, the system fetches the documents and extracts additional keywords corresponding to the selection attributes from the returned graphs which used metadata of the documents. This step ensures that the data meets the criteria specified by the user.
- Next, since the system extracts the new keywords and add them to the ware house annotating the projection attributes, later query results presented to the user for different keywords also include these documents if there is a match in the query keyword and the document content.

F. Semantic Flexibility
The use of fuzzy sets introduces two important concepts:
- Imprecision: This allows the system to handle data that might not exactly match the user's criteria but is within an acceptable range of values.
- Similarity: This ensures that the system can recognize and retrieve data that is semantically close to the user's preferences, even if not identical, but it is in different dialect.
This process enables users to perform more sophisticated and flexible queries, retrieving data that aligns closely with their needs and preferences, and thus enhancing the overall data retrieval experience.

G. Ontology-based filtering and disambiguation
1. Scoring Technique: This method boosts the accuracy of candidate concepts by checking for related concepts nearby in the text.
2. Implementation: Uses relationships between units and quantities in the ontology to validate candidate concepts, discarding those with values outside expected ranges.
3. Limitations: Less effective for large datasets and varied ontologies, as it struggles with certain values which have multiple valid interpretations.
4. Ambiguous Concepts: Current methods does not resolve ambiguities. A solution to ambiguity resolution in different dialects is required. Especially when dealing with different meaning for the same word in different dialect.



Pictures 1 and 2 : Backend cloud databases being generated by the two subsystems. Listing file source and annotations being done.

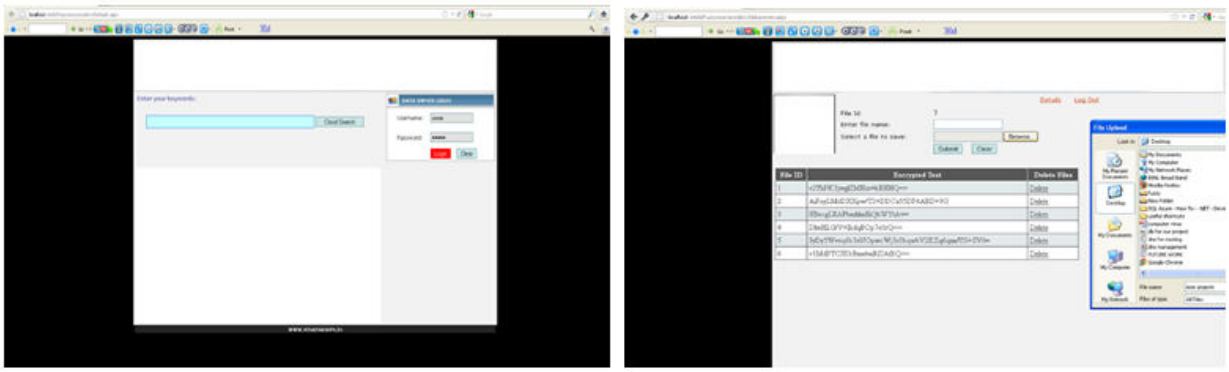
International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

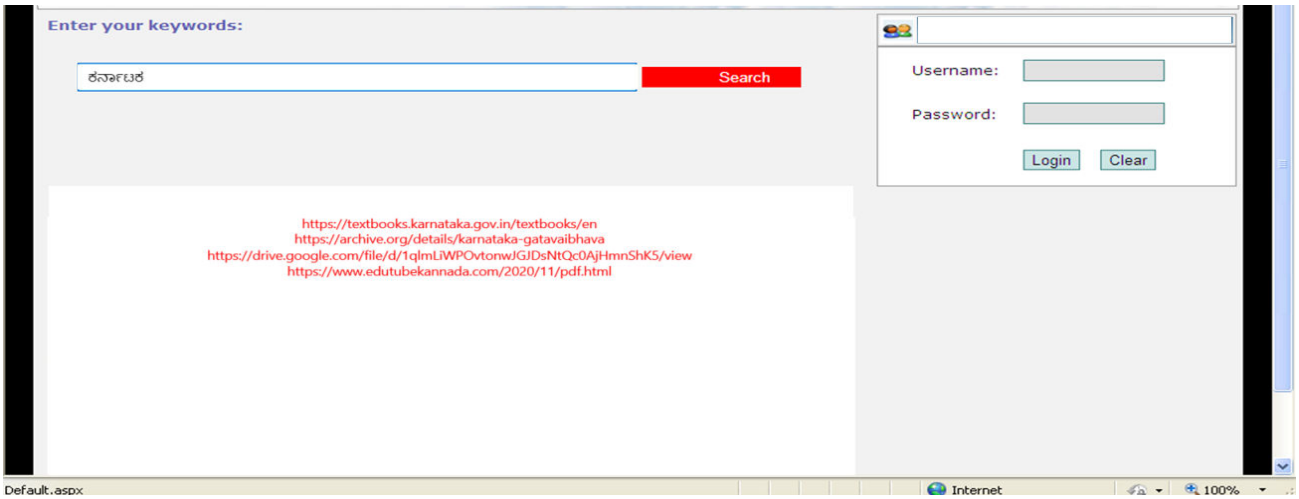
|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 11, November 2024

|DOI: 10.15680/IJIRSET.2024.1311134|



Pictures 3 and 4: Search engine and saving of file feature demonstration.



Picture 5: References shown for the keyword ಕರ್ನಾಟಕ, one of them is a google drive link having ಕರ್ನಾಟಕ.pdf for which extra keyword annotation is shown in the next picture.

Picture 6: “ಕಾಡು”, "ವನ ಪ್ರದೇಶ" and "ಅರಣ್ಯ ಪ್ರದೇಶ" annotations being added when a search for “ಕರ್ನಾಟಕ” has retrieved a document ಕರ್ನಾಟಕ.pdf for the first time from the internet using metadata ಕರ್ನಾಟಕ in the file name, its source address is added to the Terminological Resource database annotating other keywords generated using CNN on features such as horizontal and vertical histograms, and applying Fuzzy keyword similarity technique using Kannada Dictionary.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

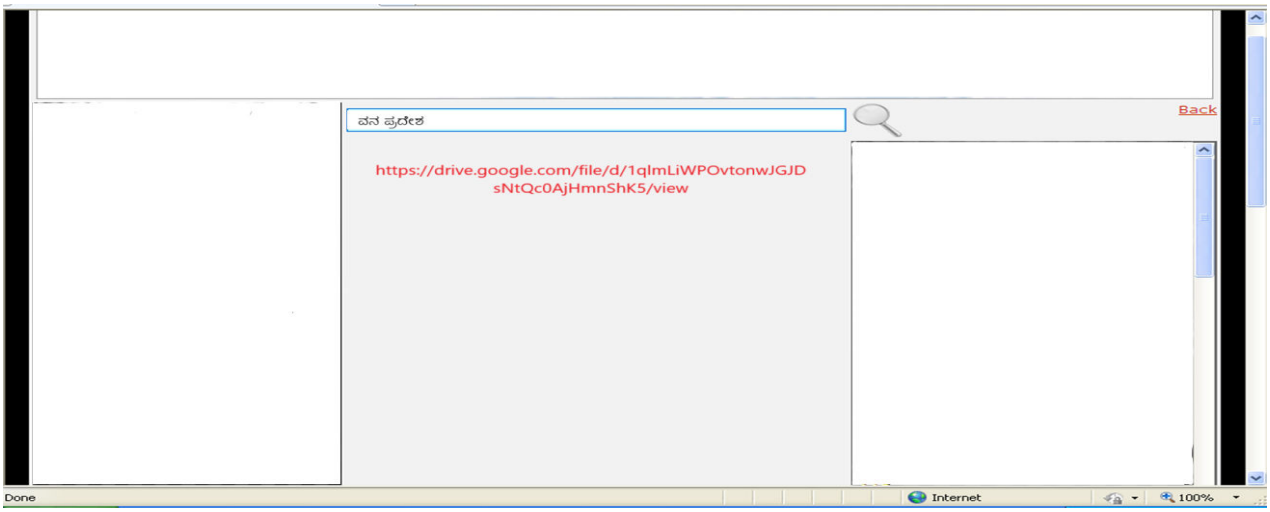
|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 11, November 2024

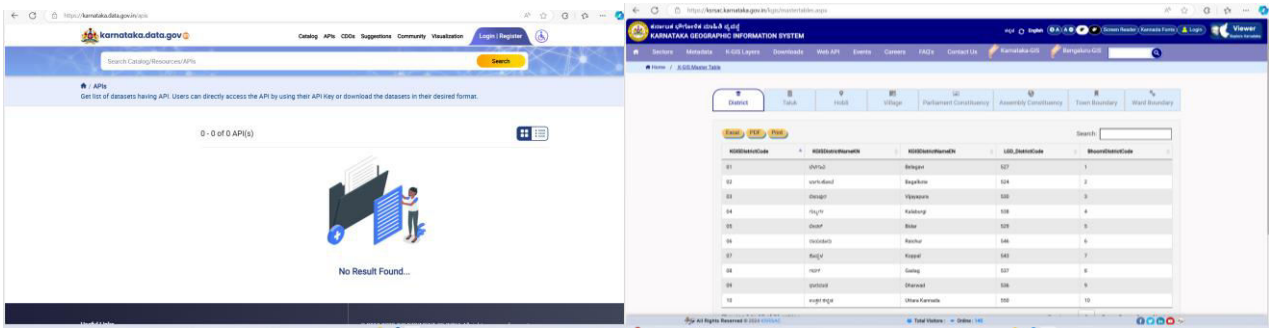
|DOI: 10.15680/IJIRSET.2024.1311134|

Then for the next time when the user searches for ವನ ಪ್ರದೇಶ the file ಕರ್ನಾಟಕ.pdf is listed even though the file meta data did not have the word phrase ವನ ಪ್ರದೇಶ. Local clustering process applied on words extracted from the document, to add the similar words of other dialects in to the Terminological Resource database, helps during query expansion.

Different document sources retrieved for the keyword ಕರ್ನಾಟಕ include a link of textbooks website of Karnataka government, Karnataka Gata Vaibhava document in archive org website, a google drive link which has a pdf document having details about Karnataka whose file name is ಕರ್ನಾಟಕ.pdf which is clicked in this scenario and a link of edutubekannada.



Picture 7: For the keyword phrase ವನ ಪ್ರದೇಶ same document reference in some google drive link ಕರ್ನಾಟಕ.pdf is listed based on the keyword extracted by the Web Text Subsystem when the keyword ಕರ್ನಾಟಕ was searched and ಕರ್ನಾಟಕ.pdf was fetched.



Pictures 8 and 9: Left picture is erroneous result when clicked on a link wrongly listed by the system. Right picture is Result showing document having ಬೆಳಗಾವಿ in a data table when the source URL was clicked for the keyword ಬೆಳಗಾವಿ.

V. CONCLUSIONS

In conclusion, the system, developed utilizing a generic Terminological Resource built using two subsystems, significantly enhances the integration and querying of XML data tables extracted from web documents and text documents. By incorporating fuzzy annotations, it enables flexible and nuanced data retrieval from different dialect documents. The system effectively annotates data tables by associating symbolic concepts, managing imprecise numerical values, and text parsing in text documents.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 11, November 2024

|DOI: 10.15680/IJIRSET.2024.1311134|

Future enhancements for the system include improving the similarity measure with additional syntactical and semantic techniques, extending semantic annotation to include textual data with different meaning in different dialects, and managing the evolution of annotation results and other ontologies. These improvements aim to further refine the system's accuracy and robustness, ensuring it remains a powerful tool for data integration and querying in diverse dialect documents.

REFERENCES

1. M. Habibi and A. Popescu-Belis, "Keyword Extraction and Clustering for Document Recommendation in Conversations" IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 23, NO. 4, APRIL 2015.
2. Patrice Buche, Juliette Dibie-Barthelemy, Liliana Ibanescu, and Lydie Soler , "Fuzzy Web Data Tables Integration Guided by an Ontological and Terminological Resource", *IEEE Transactions on Knowledge and Data Engineering*, vol.25, pp 805-819, April 2013.
3. S´ebastien Destercke and Patrice Buche and Brigitte Charnomordic, "Data reliability assessment in a data warehouse opened on the Web", *Proceedings of the 9th international conference on Flexible Query Answering Systems, Springer-Verlag Berlin*, pp 174-185, 2011.
4. P. Cimiano , P. Buitelaar , J. McCrae , M. Sintek, "LexInfo: A Declarative Model for the Lexicon-Ontology Interface", *Journal of Web Semantics*, vol. 9, no. 1, pp. 29-51, 2011.
5. John McCrae, Dennis Spohr, and Philipp Cimiano, "Linking Lexical Resources and Ontologies on the Semantic Web with lemon", *8th Extended Semantic Web Conference, The Semantic Web: Research and Applications (ESWC)*, pp. 245-259, 2011.
6. Mark van Assem, Hajo Rijgersberg, Mari Wigham and Jan Top, "Converting and Annotating Quantitative Data Tables", *Proceedings of the 9th international semantic web conference on The semantic web – Springer Volume Part I*, pp 16-31, 2010.
7. Roman Yangarber, Winston Lin, Ralph Grishman, "Unsupervised Learning of Generalized Names", *Proceedings of the 19th International conference on Computational linguistics*, vol.1 pp1-7, 2002.
8. Akshaya Zantye, V.D.Thombre, Pallavi Yevale,"OBDI System for Fuzzy Web Data Table Integration Using an Ontological and Terminological Resource", International Journal on Recent and Innovation Trends in Computing and Communication Volume: 3 Issue: 7
9. Jin Li, Qian Wang, Cong Wan, Ning Cao, Kui Ren, and Wenjing Lou, "Fuzzy Keyword Search over Encrypted Data in Cloud Computing", This paper was presented as part of the Mini-Conference at IEEE INFOCOM 2010
10. Andrew K. McCallum. 2002. MALLET: A machine learning for language toolkit. http://mallet.cs.umass.edu

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details