Information Retrieval for Marathi and English Language using Cross Temporal Information Retrieval Algorithm

Avanti Patange
Dept. of Computer Engineering
D. Y. Patil COE, Pune
patangeavanti15@gmail.com

Rupesh V. Jagtap
Dept. of Computer Engineering
D. Y. Patil COE, Pune
rupesh9028@gmail.com

Sandeep Pathak
Dept. of Computer Engineering
D. Y. Patil COE, Pune
sandepp.gate@gmail.com

***Abstract*—In this paper we propose a novel approach to convert documents written in Marathi and English vernaculars of older centuries to modern Marathi and English using cross temporal information retrieval algorithm. Researchers are developing interest in Temporal Information Retrieval as the amount of data on Internet is growing exponentially in this 21st century because of which it is difficult for the user to retrieve the relevant documents. IR research has often ignored lexical drift. But in the rising domain of huge digitized book collections, the risk of vocabulary mismatch due to language change is high. Some collections contain text written in the vernaculars of older centuries. The time dimension available in the documents should be integrated with document ranking for effective retrieval. Research in converting the vernaculars of older centuries of Marathi language to modern Marathi is missing. So we proposed a cross temporal information retrieval algorithm to solve this problem. The efficiency of our proposed system is tested on Dnyaneshwari written in 13th century books and Marathi dictionary for marathi language and books of Harry potter and English dictionary and desired results are obtained.**

***Keywords*—Temporal Information Retrieval, Cross language IR, Cross Temporal IR.**

I. Introduction

The exponential growth Internet has lead to the huge amount of data on the web, retrieving relevant data based on the user query should be the main aim of Information retrieval systems. The process of providing users with the most relevant documents from an existing collection is called as Information retrieval. User information needs are express via Queries which are mostly in short textual form. Fig. 1 shows the extensive growth of the creating, handling and sharing information on Internet.

In recent years, time has been reaching increasing grandness within search contexts, contributing to a new research area known as temporal Information retrieval (T-IR) that contains a number of different challenges. In recent year temporal information retrieval has evolved a topic of great concern. The purpose of T-IR is to mend the potency of information retrieval methods by tapping temporal information in documents and queries.

The aim of T-IR is to meet search needs by merging the traditional belief of document relevance with temporal

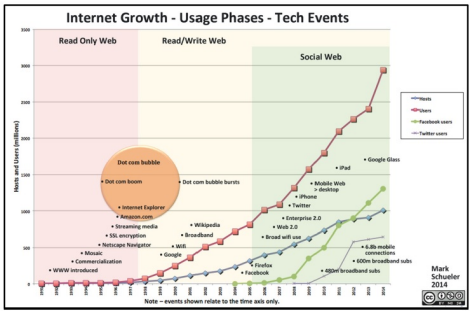

Fig. 1. Internet Growth/Usage Phases/Tech Events[1]

relevance. For example, users may ask for documents that depict the past like queries about historical image, files holding the most recent, up-to-date information like queries about weather forecasts or currency rates, or even future-related information like queries about planned events to be held in a certain area. To meet this goal these collections must support appropriate search interactions. Examples of interesting applications of temporal IR are document search, summarization, and clustering. Since the web is reliably changing, preserving up-to-date indexes is difficult task. It is hard to retrieve web documents so that their temporal dimension will satisfy the user query. Basically, two types of temporal information particularly useful for temporal IR: 1) the publication or creation time of a document, and 2) temporal expressions mentioned in a document or a query.

Information on web is growing very fast so it makes T-IR a difficult task. A clear understanding of the temporal nature of queries is hard because of query ambiguity, different possible temporal characteristics of queries and unknown user's expectations toward the required temporality of search results. Sometime returned search results do not satisfy the user information needs due to problems of a temporal nature. CTIR is needed as there is growth in number of documents in various languages over the Internet. Since queries and documents are in different languages, CTIR requires a translation phase along with the usual monolingual. Also the relevant document may

not exist in the user's native language retrieval phase, user could be able to read documents in language different from the native one, but could have difficulty in formulating queries in those languages.

Currently all the techniques are for retrieval of vernacular English documents only. Research in retrieval of Marathi vernacular documents is still missing. In this paper we give a brief idea of retrieval of Marathi vernaculars documents and also English vernacular documents retrieval. There are many books which are written in vernacular English and Marathi language. So there is a challenge to retrieve the documents based on distribution of query words observed in documents. Our proposed architecture focus on the retrieval of Marathi vernacular document based on the query given by the user and converting those Marathi and English vernacular to modern Marathi and English language. Thus a person interested in Marathi or English documents will be able to find liberal range of documents from a variety of historical periods using single query. It is somehow same as cross language information retrieval (CLIR).

To test the efficiency of the proposes system Dnyaneshwari, Harry Potter books and English,Marathi dictionary are used as the dataset.

II. Related Study

Historical linguists have lent sophistication and rigor to the study of language change. Most notable in our context is Edward Sapir's theories on language drift, the process by which a language changes over time to such an extent those earlier texts becomes unintelligible to contemporary readers [2]. However, although historical linguistics could surely help with the problem of CTIR, its focus on the analysis of linguistic genealogies is only indirectly useful here.

In domains closer to IR, work on document date identification relates to our study. In [3], [4], and [5], variations on language modeling have been brought to bear on the problem of identifying when a given document was written. As in these studies, our approaches to CTIR are based on language models, resting on the assumption that resolving differences among distant centuries vernaculars are tractable via a high-level estimation of word probabilities.

From a conceptual and practical standpoint, CTIR is similar to two active IR research areas: cross-language IR and IR over noisy text (typically due to OCR errors). If we consider archaic English and Modern English to be two distinct languages, then the CTIR problem as we have presented it is a cross-lingual retrieval problem. Certainly, CLIR methods as surveyed in [6] have a role to play in CTIR. For instance, reliance on dictionary-based and corpus-based translation tools enter into our analysis. Additionally, the approach to structuring queries that we adopt is common in the CLIR literature [7]. Efforts to solve the vocabulary mismatch problem have brought CLIR methods into monolingual retrieval, as well. Work such as [8] and [9] considers query-document matching as an English-to-English translation process.

The CTIR problem also echoes challenges faced by retrieval over modern OCR'd text. Of course, much of the research on IR for scanned documents is situated in the domain of cultural heritage [10], as is our work. But the similarity is more substantive that this. Foundational work such as [11] presents n-gram methods for supporting search over degraded texts. Of particular interest is [12], where Lam-Adesina note that many IR methods are quite robust against OCR-introduced noise but relevance feedback is more brittle.

Finally, the task of CTIR is not constrained to search over digitized books. But these collections do make the problem of language drift particularly keen. A good deal of work in the INEX book track has treated OCR-related problems in book search [13]. However, the focus of INEX topics has left linguistic change as a relatively minor problem in most previous studies of book search. Perhaps the most important fact about the literature on book search is that while digitized book data is plentiful, what people will want to do with these data is still largely unknown.

The existing system studied are only for vernacular English language only, no study is yet done on retrieval of Marathi vernacular this motivated us to take step forward to convert the Marathi vernacular to modern Marathi.

III. Applications of Temporal Information Retrieval

There are various applications of temporal information retrieval system.

A. Existing Temporal Search Engines

An enormous amount of information is stored in web archives including web pages harvested and added into the archive repository when recrawling, as well as focused web warehouses like news archives. Information in such document repositories are helpful for both skilled users, e.g., historians, librarians, and journalists, as well as a general user searching for information needs in old versions of web pages. To date, there are existing search systems that provide accessibility to web archives. For example Google News Archive Search. This tool allows a user to search a news archive using a keyword query and a date range. In addition, the tool provides the ability to rank search results by relevance or date. However, there is a problem that has not been addressed by this tool yet, e.g., the effect of terminology evolution. Consider the example; a user wants to search for news about Earthquake in Maharashtra that are written before 2000. So, the user issues the query Earthquake in Maharashtra and specifies the temporal criteria 1990/01/01 to 2000/31/12. So only a small number of documents are returned by the tool where most of them are not relevant to the Earthquake in Maharashtra. In other words, this problem can be viewed as vocabulary mismatch caused by the fact that the term Earthquake in Maharashtra was not widely used.

B. Analysis and Exploration over Time

Some applications have taken time-based exploration of textual archives beyond just searching over time. Filtering

and displaying information might benefit from presenting time information conveniently in some domains.[14] Time Explorer combines a number of interesting features present in other time-based systems, although extended in several important ways. First, users are enticed to discover how entities such as people and locations related with a query change over time. Second, by seeking on time expressions elicited automatically from text, the application allows the user to research not only how topics developed in the past, but also how they will continue to evolve in the future. All these features are combined in an intuitive easy-to-use interface, which is always a great challenge when designing search engines that allow for extended capabilities. [15]Other exploration-based systems have turned into other sources of textual information, for instance word evolution over time. This is naturally promising research direction, since there are available digitalized collections that span centuries. These systems need to employ a combination of several time-aware algorithms. For instance, they will require to extract time-related information from a given underlying textual collection, index this information along with the standard collection's contents and perform a combination of temporal query analysis and ranking after a user query is posed.

C. Temporal Summarization

Another stream of applications has focused into exploiting the use of time for enhanced story telling. [16]The task of news summarization has been studied previously ranging from multi-document summarization to generate a time-line summary for a specific news story. A user enters a topic into a news search engine and obtains a list of relevant results, ordered by time. [17]Furthermore, the user subscribes to this query so in the future she will continue to receive the latest news on this query. The time dimension comes into play when the user is observing a current document, and one may want to show the most relevant entities of the document for her query taking into account features extracted from previous documents. [18]The most widespread summarization technology is the focused summaries produced by search engines, or search results snippets. Those are useful to assist users in deciding whether a document is relevant for a query or not.

D. Temporal Clustering of Search Results

Another popular application that makes use of time-based IR techniques is search results clustering, which is an important feature for some information retrieval applications, in particular, enterprise search systems. [19]There is a prototype that is able to display date and time attributes per cluster. Those attributes are extracted from the textual content of documents that belong to a particular cluster. Furthermore, [20] extend the idea of reusing temporal information embedded in documents to enhance results presentation by introducing a time-line-based display of results. The time lines span different time granularities and display temporal information extracted automatically from documents and made explicit. They also explore how search results can be clustered according to time and how to produce temporal snippets to navigate through documents returned. [21]Describe a method to group search result documents at a year level using a similarity measure that identifies the most relevant dates. Each group is then displayed differently in the results page, allowing for an easier exploration of the search results, as demonstrated by a user survey.

IV. Proposed System

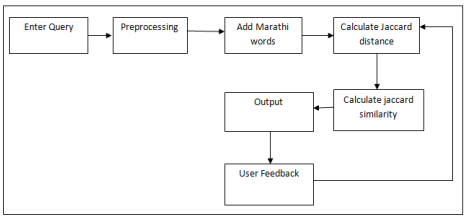

Fig. 2. System Architecture

Fig. 2 shows the architecture of the proposed system which aims at converting vernacular Marathi to modern Marathi When user enters a query data preprocessing is done. That is Marathi words get added from the dictionary. Jaccard similarity is calculated between user string and dictionary string as

$$J(x,y) = \frac{\sum_i min(x_i, y_i)}{\sum_i max(x_i, y_i)} \quad (1)$$

(where x and y are set of document) and also Jaccard distance is calculated as

$$d_i(x,y) = 1 - J(x,y) \quad (2)$$

based on which result is retrieved. If user is not satisfied with the retrieved result then user can give his feedback and again same procedure is followed and result is given to the user.

We have used Harry Potter story book written in English which contain 200 chapters and Dnyaneshwari book written in historical Marathi language which contain 250 stories as a dataset. These datasets are gathered from Google books.

A. Cross Temporal Information Retrieval Algorithm

1) Create database of word and rules.
2) Normalize rules in Marathi.
3) Check similarity between input word and database & add word (dictionary of Marathi word) using Jaccard similarity.
4) Find semantics similarity of given word.
5) Get feedback of user.
6) Calculate posterior probability, rank them next time.
7) End

V. Results

Fig. 3 projects the comparison graph between existing system and proposed system.

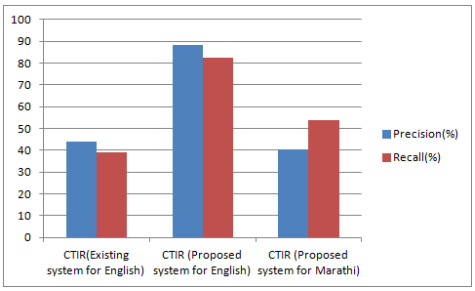

Fig. 3. Performance Graph

TABLE I
PERFORMANCE EVALUATION

System	CTIR (Existing system for English)	CTIR (Proposed system for English)	CTIR (Proposed system for Marathi)
Precision (%)	43.8	88.4	40.4
Recall (%)	39.2	82.6	53.7

These are observed results for precision and recall graph calculated from the formula as follows:-

Precision = No. of relevant documents retrieved / Total no. of documents retrieved

Recall = No. of relevant document retrieved / Total no. of documents in Database

The results generate in Table 1, precision and recall of proposed English system is high as compare to existing system for English language. And for CTIR system for Marathi language the precision is 40.4% &recall is 53.7% which shows that the proposed system results are improvised and the combine approach has enhanced the TIR process.

VI. CONCLUSION

In this paper we have proposed a system which combinely deal with both Marathi and English vernaculars and convert them t respective modern languages. This work is novel for Marathi conversion. Dnyaneshwari along with Marathi dictionary are used to test Marathi system Performance. English vernacular based system is tested with standered English dictionary and old Harry Potter books. Comparative study of existing English language system shows the better performance of later one.

REFERENCES

[1] Alonso, Omar Rogelio. “Temporal information retrieval.” Computer Science (2008): 1-155.
[2] Edward Sapir. Language: And Introduction to the Study of Speech. Harcourt, Brace, New York, 1921.
[3] Kanhabua, Nattiya, and Kjetil Nrvg. “Improving temporal language models for determining time of non-timestamped documents.” Research and advanced technology for digital libraries. Springer Berlin Heidelberg, 2008. 358-370.
[4] Kumar, Abhimanu, Matthew Lease, and Jason Baldridge. “Supervised language modeling for temporal resolution of texts.” Proceedings of the 20th ACM international conference on Information and knowledge management. ACM, 2011.
[5] Jong, de FMG, Henning Rode, and Djoerd Hiemstra. “”Temporal language models for the disclosure of historical text.” Royal Netherlands Academy of Arts and Sciences, 2005.
[6] Jurafsky, Dan, and James H. Martin. Speech & language processing. Pearson Education India, 2000.
[7] Darwish, Kareem, and Douglas W. Oard. ”“Probabilistic structured query methods.” Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. ACM, 2003.
[8] Berger, Adam, and John Lafferty. “”Information retrieval as statistical translation.” Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1999.
[9] Karimzadehgan, Maryam, and ChengXiang Zhai. “”Estimation of statistical translation models based on mutual information for ad hoc information retrieval.” Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval. ACM, 2010.
[10] Droettboom, Michael. “”Correcting broken characters in the recognition of historical printed documents.” Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries. IEEE Computer Society, 2003.
[11] Harding, Stephen M., W. Bruce Croft, and C. Weir. ”Probabilistic retrieval of ocr degraded text using n-grams.” Research and advanced technology for digital libraries. Springer Berlin Heidelberg, 1997. 345-359.
[12] Lam-Adesina, Adenike M., and Gareth JF Jones. “”Examining and improving the effectiveness of relevance feedback for retrieval of scanned text documents.” Information processing & management 42.3 (2006): 633-649.
[13] Kazai, Gabriella, and Antoine Doucet. “”Overview of the INEX 2007 book search track: BookSearch’07.” ACM SIGIR Forum. Vol. 42. No. 1. ACM, 2008.
[14] Koen, Douglas B., and Walter Bender. “Time frames: Temporal augmentation of the news.” IBM systems journal 39, no. 3.4 (2000): 597-616.
[15] Matthews, Michael, Pancho Tolchinsky, Roi Blanco, Jordi Atserias, Peter Mika, and Hugo Zaragoza. “Searching through time in the New York Times.” In Proc. of the 4th Workshop on Human-Computer Interaction and Information Retrieval, pp. 41-44. 2010.
[16] Erkan, Gnes, and Dragomir R. Radev. “LexRank: Graph-based lexical centrality as salience in text summarization.” Journal of Artificial Intelligence Research (2004): 457-479.
[17] Yan, Rui, Xiaojun Wan, Jahna Otterbacher, Liang Kong, Xiaoming Li, and Yan Zhang. “Evolutionary timeline summarization: a balanced optimization framework via iterative substitution.” In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, pp. 745-754. ACM, 2011.
[18] Sipos, Ruben, Adith Swaminathan, Pannaga Shivaswamy, and Thorsten Joachims. “Temporal corpus summarization using submodular word coverage.” In Proceedings of the 21st ACM international conference on Information and knowledge management, pp. 754-763. ACM, 2012.
[19] Alonso, Omar, and Michael Gertz. “Clustering of search results using temporal attributes.” In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 597-598. ACM, 2006.
[20] Alonso, Omar, Michael Gertz, and Ricardo Baeza-Yates. “Clustering and exploring search results using timeline constructions.” In Proceedings of the 18th ACM conference on Information and knowledge management, pp. 97-106. ACM, 2009.
[21] Campos, Rui, Alipio Mario Jorge, Guilherme Dias, and Celia Nunes. “Disambiguating implicit temporal queries by clustering top relevant dates in web snippets.” In Web Intelligence and Intelligent Agent Technology (WI-IAT), 2012 IEEE/WIC/ACM International Conferences on, vol. 1, pp. 1-8. IEEE, 2012.