# Investigation of the Issues Related to Word Level Features of Marathi Language for Searching Web Content

Harshali B. Patil
School of Computer Sciences
North Maharashtra University
Jalgaon- Maharashtra

Ajay S. Patil
School of Computer Sciences
North Maharashtra University
Jalgaon- Maharashtra

## ABSTRACT
Now-a-days Internet is an important source of information. So everyone uses Internet in their day-to-day life for doing various activities. Text search is one of the important activities performed daily by the Internet users. The dramatic growth in textual data available on Internet in regional languages gave birth to natural language search. Natural language content retrieval poses certain problems due to the word level features of that language such as spelling variation, morphological variation, etc. The unavailability of tools and techniques for these regional languages can be the reason for the low recall level for these natural languages information retrieval. This paper addresses the issues related to Marathi language word level features for textual content retrieval. This paper describes different types of problems with examples and also suggests solutions to these problems.

## General Terms
Information retrieval, natural language processing.

## Keywords
Natural language processing, text retrieval, morphology, Marathi

## 1. INTRODUCTION
The development of the World Wide Web (WWW) and its usage during last 10-15 years has both increased the number of digital documents and the variety of languages appearing on the web. In early days of the Web, its monolingual nature concentrated the research in IR for English language only. Since the Web became multilingual, the content on the Web started appearing in regional languages of the world over and is continuously increasing day by day. Retrieving the meaningful information from this enormous e-data available is a challenging task. The work related to Indian languages information retrieval has been recently started. Due to the increasing proliferation of the Internet in India there is a dramatic growth in textual content available in Indic languages on World Wide Web. Table 1 describes the information related to the scenario of Internet users in top 5 countries. From table 1 it is observed that India is on second position in the list of countries according to number of Internet users. India shares 13.5 % of total users all over the world and tremendous increase in number of Internet users is observed from figure 1. It shows the growth of Internet users in India. Since last two years near about one hundred million users are increased.

**Table 1 : Top 5 countries by Internet usage till June 2016** [+]

| Rank | Country | Internet users | Growth (*) 2000 – 2016 (%) | Country's share of world's Internet user |
|---|---|---|---|---|
| 1 | China | 721,434,547 | 3,106.4 | 21.1 % |
| 2 | India | 462,124,989 | 9,142.5 | 13.5 % |
| 3 | United States | 286,942,362 | 200.9 | 8.4 % |
| 4 | Brazil | 139,111,185 | 2,682.2 | 4.1 % |
| 5 | Japan | 115,111,595 | 144.5 | 3.4 % |

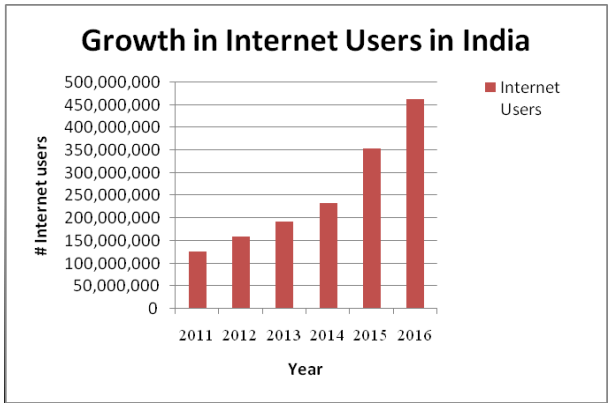


**Fig. 1: Internet growth in India** [*]

India is a multilingual country with more than 20 official languages. The Internet users in India either use English or the local languages. Following table shows the distribution of the Internet users in India.

**Table 2: Internet users in India**

| Region | Internet users(Mn) | Local language users (Mn) |
|---|---|---|
| Urban | 188 | 81 |
| Rural | 81 | 46 |
| Total | 269 | 127 |

From table 2 it is observed that near about 47% users are local language Internet users among total Internet users. Table 3 presents the information about some of the Indian local languages used online.

**Table 3: Local languages used online[1]**

| Language | Rural (%) | Urban (%) |
|---|---|---|
| Hindi | 27 | 60 |
| Marathi | 19 | 10 |
| Tamil | 12 | 20 |

Form table 3 is observed that Marathi is one of the important languages used in both rural as well as urban segment. It is an important language in India. Marathi is one of the official languages ranking fourth in number of speakers in India [1]. It is the official and regional language of state of Maharashtra and Goa. Total 627 websites are hosted in Marathi [2]. In rural India, 19 percent of the users use Marathi while in urban India 10 percent of the user uses it online.

E-mail, news and search engines are the popular applications that are used by local language users in urban segment[1]. Searching is finding the relevant information according to users need. This need for information is mentioned in various forms of queries like keyword queries, phrase queries, etc. Many users want to search the information in their local language. Basically queries are composed of words. Words are made up of more basic units called as morphemes which can be defined as smallest part of word that has grammatical function or meaning. In linguistics there are many words related to the same morpheme which represents the same concept. The rapid growth of web content available in regional languages attracted the researchers to develop tools and techniques for efficient information retrieval in Indian languages. The table 4 shows that the amount of Indian language web content available in electronic form has reached to the significant stage.

**Table 4: Hits generated for Marathi terms**

| Search engine→ Search term ↓ | Google | Bing | Khoj |
|---|---|---|---|
| चेन्नई एक्सप्रेस | 334,000 | 13,700 | 15,300 |
| आशिकी 2 | 278,000 | 27,100 | 18100 |
| गेट 2014 | 2,620,000 | 23,100 | 33500 |
| अब्दुल कलाम | 381,000 | 22,000 | 37400 |

Many newspapers such as Lokmat, Daily Sakal, Loksatta, Divya Marathi etc. have online editions available in Marathi. Besides these there are various magazines and information portals serving content in Marathi language. Marathi is used as an official language of Maharashtra and Goa states. Many government sites such as [https://www.maharashtra.gov.in] are available in Marathi language and contain huge amount of important data. With all this information available online, there is a growing need for techniques to effectively manage the available information. Unfortunately, in many cases, directly using the existing search systems on Marathi text collections does not yield very good results. This paper highlights the issues related to Marathi text content retrieval.

## 2. RELATED WORK

Strzalkowski et. al studied information retrieval using robust natural language processing and pointed a feasible direction for integrating NLP into the traditional IR [2]. Brants reviewed NLP techniques in IR and saw a modest benefit of NLP techniques in IR [3]. Majumder et.al. reviewed the current state of the art in mono-lingual and cross- lingual information access in Indian languages and outlines the project which aims to develop a IR system for Indian languages [4]. Reductive and generative approaches to morphological variation of keywords in monolingual information retrieval has been studied by Kettunen where he found that stemming, inflectional stem generation and its enhancements and most frequent case form generation of keywords gave good retrieval results when compared to lemmatization [5]. Pal et. al studied issues in searching for Indian language web content [6]. Soundalgekar reported the work related to Internet search for Indian languages where he suggested that for languages other than English , it may be useful or needful to make use of stemming[7]. The working of stemmer and comprehensive analysis of stemmers available for Indian languages had been discussed by Patil et al. [8,9]. Patil el al carried out the work related to Marathi link grammar parser [10].

## 3. EXPERIMENTAL SETUP

The word level features of Marathi language may create lots of problems in IR. This work reports the deviation occurred in search due to the word level features of Marathi. The word level features discussed here include spelling variation, inflection, derivation, compound word and phrases, ambiguity, and synonyms related to Marathi words. The most popular Google search engine has been used for searching the information and the hits generated contains the pages in Marathi as well as other languages like Hindi due to the use of exact similar word and similar script in this language.

### 3.1 Spelling Variation

Spelling variation or change in surface form may cause problems for searching. In Indic languages the spelling variation occurs due to the existence of (i) long and short forms of certain vowels, (ii) sibilants and (iii) several forms of nasal sound [6] and the same is applicable for Marathi language. According to Patil et.al lots of variations exists in spellings writing style [11,12]. The 'श' character is written in two different ways, some words which include this character are written as प्रश्न or प्रश्न, विश्व or विश्व. Long or short vowels are used interchangeably. In Marathi language if the syllable formed includes two consonants in the sequence among which the first consonant is not a complete consonant i.e., if it does not have any vowels then two different methods of writing exist and used interchangeably, so the same words can be written in two different ways Eg. h_ or हद्द when different fonts are used. Several surface forms are found for the words which includes nasal sound in Marathi words. The nasal sound in Marathi language is denoted by using these " ं, ङ्,ञ्, ण्, न्, म् " characters. Ex- हिन्दी or हिंदी. Change in surface forms may also occurs due to orthographic rules for the language when there has been orthographic reform in the language. Some examples of spelling variations along with hits generated while searching using Google search engine are presented in table 5.

**Table 5 : Hits generated for same word's different spelling**

| Variation 1 | Hits generated | Variation 2 | Hits generated |
|---|---|---|---|
| सिता | 704,000 | सीता | 2,280,000 |
| अश्व | 233,000 | अश्व | 1,110 |
| कण्हेरी | 9,150 | कन्हेरी | 6,310 |

| | | | |
|---|---|---|---|
| सुध्दा | 247,000 | सुद्धा | 367,000 |
| तु | 3,670,000 | तू | 8,740,000 |
| शितल | 385,000 | शीतल | 1,300,000 |

## 3.2 Inflections

Marathi is a morphologically rich language. Singular and plural nominal forms are different in Marathi language. Ex- माळ- माळा , भिंत- भिंती , नदी- नद्या, कुत्रा- कुत्रे. Inflectional morphemes do not create separate words they only modify the words in which they occur in order to indicate grammatical properties such as plurality or past tense. Eight grammatical cases exists in Marathi language. When any case marker is attached to the root/stem the original form of that root/stem is first changed and then the case marker is attached with it. Ex- कुत्रा (कुत्र्या ) + ला = कुत्र्याला. For denoting the relation of the term with other terms like verb in that sentence the case markers are attached to that term causes inflection ex- रामाने रावणास मारले. The Marathi verbs inflect for person number, gender, tense, aspect and mode. The different affixes modify the meaning of the root and may hide it in retrieval. Table 6 shows the hits generated for different inflections of the same word.

## 3.3 Derivations

Derivation is the process of creating separate words which are morphologically related with each other. It involves one or more changes in the form. Sometimes the words are derived either by attaching suffix (हुशार – हुशारपणा) or by adding prefix (प्रत्यक्ष – अप्रत्यक्ष , उत्साही – निरुत्साही ). These types of words should sometimes be conflated in IR but which sometimes have been lexicalized to the degree that the semantic collection to the root is only formal.

**Table 6: Hits generated for the word and the corresponding derived word**

| Word | POS | Hits generated | Derived word | POS | Hits generated |
|---|---|---|---|---|---|
| हुशार | Adj | 507,000 | हुशारपणा | N | 442 |
| एक | Adj | 122,000,000 | एकदा | Adv | 4,490,000 |
| कोणी | P | 3,220,000 | कोणीकडून | Adv | 703 |
| जो | P | 39,400,000 | जोपर्यंत | Adv | 348,000 |
| त्या | P | 11,800,000 | त्याअर्थी | Adv | 20,300 |
| दिवस | N | 17,800,000 | दिवसा | Adv | 359,000 |
| घर | N | 48,100,000 | घराचा | Adj | 148,000 |
| समोर | PP | 3,610,000 | समोरचा | Adj | 97,900 |

Table 6 shows the deviation occurred in the word and its derived word in terms of number of hits generated while searching on the web using Google search engine. Here Adj represents adjective, N represents noun, P for pronoun, PP denotes preposition and Adv represents adverb. Table 7 clearly shows that very much deviation exists for searching the derived word form as compared to root word. This information motivates for the development and use of word normalization tools like stemming, lemmatization or morphological analyzer for IR.

**Table 7: Hits generated for different inflections of the same word**

| Word | Hits | Variation 1 | Hits | Variation 2 | Hits | Variation 3 | Hits |
|---|---|---|---|---|---|---|---|
| माळ | 340,000 | माळा | 125,000 | माळेतील | 5,080 | माळांचा | 1,660 |
| भिंत | 477,000 | भिंती | 244,000 | भिंतीला | 61,300 | भिंतीवर | 147,000 |
| नदी | 6,980,000 | नद्या | 215,000 | नद | 182,000 | नद्यांचा | 26,800 |
| कुत्रा | 427,000 | कुत्रे | 400,000 | कुत्र्याला | 78,000 | कुत्रं | 19,700 |
| मी | 13,600,000 | आम्ही | 7,070,000 | आपण | 9,720,000 | आम्हांस | 46,800 |
| हा | 17,900,000 | ह्या | 4,640,000 | हे | 28,600,000 | ही | 68,700,000 |
| ये | 44,200,000 | या | 66,500,000 | याल | 96,400 | येशील | 283,000 |
| गाणे | 947,000 | गाऊ | 259,000 | गाइले | 15,900 | गातोस | 1,800 |
| बस | 20,900,000 | बसा | 744,000 | बसतील | 67,900 | बसेन | 49,200 |
| भारत | 57,500,000 | भारताचा | 1,040,000 | भारताने | 649,000 | भारतावर | 233,000 |
| नासा | 949,000 | नासाची | 4,240 | नासाने | 13,900 | नासावर | 114 |
| घर | 48,100,000 | घरासमोरचा | 2,760 | घराकडे | 115,000 | घरातून | 349,000 |

## 3.4 Compound words and phrases

The words which contain two or more root words out of which all, one, or none of them may be bounded are called as compounds ex- नवी दिल्ली. In compound words generally one of them is the head word and other are its modifiers. Modifiers are generally written before the head word. Compound words can be written as various ways in Marathi language. Sometimes they are written as single word like नगररचना, sometimes these words and phrases parts are connected by hyphen like एक-दोन or sometimes they are written as two different words as in नवी दिल्ली . For representing compound words there is often instability in surface expression and many times compounds and phrases carry meaning that is more than product of the meaning of their constituents i.e they are lexicalized in their meaning. When the compounds are written together their headwords may be inaccessible in retrieval. The following table presents the difference between the numbers of hits generated for compound words if they are written in different ways.

**Table 8: Hits generated for variation of compound words**

| Compound word | Variation1 | Variation 2 |
|---|---|---|
| नवी दिल्ली | 332,000 | 7,480 |
| काळा पैसा | 155,000 | 11,800 |
| सुख दुःख | 195,000 | 43,800 |
| हास्य प्रधान | 2,070 | 6,980 |
| बरे वाईट | 26,000 | 15,600 |
| दिवस रात्र | 140,000 | 70,700 |
| ये जा | 132,000 | 7,690 |
| अणु रेणू | 10,900 | 1,480 |
| गाठ भेट | 32,700 | 4,920 |
| उलट्या सुलट्या | 832 | 88 |

## 3.5 Ambiguity

Natural language is ambiguous because there are many meanings associated with the same word or different words exist for the same meaning. Determining the correct sense to the term is easy for humans but ambiguity determination and resolution is a challenging task for computerized language processing. Marathi is highly ambiguous language. Here the different senses of the same word in top 10 results were observed for determining the ambiguity level present in textual web content search for Marathi language.

**Table 9: Variation in top 10 results of searching for ambiguous words according to different sense**

| Word | Sense 1 | Sense 2 | Sense 3 |
|---|---|---|---|
| कर | Do(03) | Tax(07) | Hand(00) |
| झाड | Tree (08) | To sweep (02) | - |
| रस | Juice (04) | *Rus* in life (06) | Interest (00) |
| घन | Cloud (02) | Cube (04) | Power ($2^3$) 04) |
| काळ | Tense (02) | Time (08) | - |

## 3.6 Synonyms

Many synonymous expressions are exits for many concepts in Marathi. Acronyms, abbreviations, antonyms can also be considered as special cases of synonymy. Paraphrasing may also be used in absence of specific word. This leads to situations where queries and documents may use different words for the same concept. Here an attempt has been made for checking the distinction obtained in results in terms of hits generated for searching the term and its abbreviation or synonym on web.

**Table 10: Hits generated for different synonyms**

| Term | Hits generated | Synonyms / Abbreviations | Hits generated |
|---|---|---|---|
| आई | 22,900,000 | माता | 9,830,000 |
| | | जननी | 659,000 |
| | | माय | 1,250,000 |
| दिन | 49,200,000 | दिवस | 17,800,000 |
| नित्य | 999,000 | नेहमी | 2,790,000 |
| | | सदा | 3,000,000 |
| वारंवार | 1,110,000 | पुन्हापुन्हा | 17,000 |
| | | पुनः पुनः | 35,800 |
| दररोज | 1,070,000 | प्रतिदिन | 1,820,000 |
| सुरेख | 363,000 | सुंदर | 10,500,000 |
| | | चांगल | 165,000 |
| उदाहरण | 4,530,000 | उदा. | 692,000 |
| महाराष्ट्र टाइम्स | 1,580,000 | मटा | 773,000 |
| प्राध्यापक | 1,020,000 | प्रा | 4,770,000 |
| फेब्रुवारी | 2,570,000 | फेब्रु | 146,000 |

## 4. CONCLUSION

It has been observed that if the inflected form of the word is used for searching the web content then the retrieved results are less. It may be the reason for low level of precision and recall. The use of original word form instead of inflected form we got more number of results so precision and recall levels can be more, so it can be concluded that word normalization process like stemming in necessary to increase effectiveness of Marathi textual search. By considering the large number of documents present on web and the large number of search query submitted to search engine for web search the normalization scheme used is required to be simple and efficient for using it in practice. A simple stemmer could be constructed by creating a list of most commonly used suffixes and deleting matching suffix from the ends of the words. The dictionary which contains synonyms can be also useful for improving the results. In future the authors intend to develop a stemmer for Marathi language and evaluate the impact of that stemmer for searching Marathi content.

## 5. REFERENCES

[1] Mhaske N, and Patil A. 2016. Issues and Challenges in Analyzing Opinions in Marathi Text. International Journal of Computer Science Issues, Volume 13, Issue 2, pp- 19-25.

[2] Strzalkowski, T. and Vauthey, B. 1992. Information Retrieval Using Robust Natural Language Processing. In Proceedings of ACL-92, pp 104–111, Newark, Deleware, USA.

[3] Brants T. 2003. Natural Language Processing in Information Retrieval, 14th meeting of computational linguistics in the Netherlands.

[4] Majumder P., Mitra M. 2009. Indian Language Information Retrieval. Guide to OCR for Indic Scripts, pp 301-314.

[5] Kimmo Kettunen. 2007. Reductive and Generative Approaches to Morphological Variation of Keywords in Monolingual Information Retrieval, Doctoral Thesis. University of Tampere.

[6] Pal D., Majumder P., Mitra M., Mitra S., and Sen A. 2008. Issues in Searching for Indian language Web Content. In iNEWS '08 Proceedings of the 2nd ACM Workshop on Improving non-English Web Searching Pages 93-96.

[7] Soundalgekar M. Internet search for Indian Languages. M.Tech dissertation, IIT Bombay.

[8] Patil H. B., Pawar B. V., Patil A. S., 2016. A Comprehensive Analysis of Stemmers Available for Indic Languages. International Journal on Natural Language Computing (IJNLC) Volume 5 – No.1, pp 45-55.

[9] Patil H. B., Patil A. S., Pawar B. V., 2014. Part-of-Speech Tagger for Marathi Language using Limited Training Corpora. IJCA Proceedings on National Conference on Recent Advances in Information Technology NCRAIT(4), 2014, pp. 33-37.

[10] Vaishali. B. Patil & B. V. Pawar 2015. Modeling Complex Sentences for Parsing through Marathi Link Grammar Parser, Int. Journal of Computer Science Issues, Vol. 12, Issue 1, No. 2, pp 108-113.

[11] Patil N. V., Patil A. S., Pawar B. V. 2016. Survey of Named Entity Recognition Systems with respect to Indian and Foreign languages. International Journal of Computer Applications (0975 – 8887) Volume 134 – No.16, pp 21-26.

[12] Patil N. V., Patil A. S., Pawar B. V. 2016. Issues and Challenges in Marathi Named Recognition. International Journal on Natural Language Computing (IJNLC) Volume 5 – No.1, pp 15-30.

[13] Feldman S. 1999. NLP meets the Jabberwocky: Natural Language Processing in Information Retrieval, http://www.scism.lsbu.ac.uk/inmandw/ir/jaberwocky.htm

[14] V, Mari and Pedraza-Jimenez 2007, Rafael Natural Language Processing in Textual Information Retrieval and Related Topics. Hipertext.net, n. 5.

[15] Govilkar L. Marathiche Wyakaran, Mehata Publishing House.

[16] Bhagwat S. Tumche Aamche Marathi Vyakaran Vidyabharati Prakashan