

Query Optimization

Amit Jena
(A115002)

Guided By
Dr. Rakesh Chandra Balabantaray

M.Tech (2nd Year)
Department of Computer Science and Engineering
IIIT, Bhubaneswar

May 22, 2017



Overview

- 1 Introduction
 - Problem Statement
 - Motivation
- 2 Literature Survey
- 3 Model
 - Module I : Conversion to Standard English Query
 - Module II : Query Sequence analyser
 - Module III : Implementation in Solr
- 4 Results



Problem statement

“Query refinement using **Query sequence** for Hindi-English
Code-Mixed Query.”

- **Code-Mixed** : The search query has a mix of Hindi and English words written in **Roman** script.
- **Query sequence** : The current query will be refined on the basis of previous queries.



Motivation :

- To fetch highly relevant information.
- Consider the user intent while optimizing query.
- Show the user what they might LIKE to see.
- Higher user satisfaction.
- More relevant results.



II. Literature Survey

Understanding “How Search Works: From Algorithms to Answers!!”

- Crawling and Indexing
 - Search engine sort the pages by their content and other factors.
 - And they keep track of it all in the index.
 - Search starts with the Web and it's made up of over 130 trillion individual pages.
- Algorithms
 - As we search, algorithms get to work looking for clues to better understand what we meant.
 - Based on these clues relevant documents are fetched from the index.
 - Then the results are ranked based on over 200 factors.



Research Papers

Title 1. Reprint of: The anatomy of a large-scale hypertextual web search engine.[8]

Key Observations :

- This paper addresses the question of how to build a practical large scale system which can exploit the additional information present in hypertext.
- They also addressed the issue of how to effectively deal with uncontrolled hypertext collection, where anyone can publish anything they want.
- Anchor text: Most search engines associate the text of a link with the page that the link is on. Here they associate it with the page the link point to.



Research Papers

Title 2. A new algorithm for inferring user search goals with feedback sessions.[9]

Key Observations :

- They proposed a framework to discover different user search goals for a query by clustering the proposed feedback sessions.
- The generated pseudo-documents to better represent the feedback sessions for clustering.
- Finally, they proposed a new criterion “Classified Average Precision (CAP)” to evaluate the performance of inferring user search goals.



Research Papers

Title 3. Lessons from the journey: a query log analysis of within-session learning.[10]

Key Observations :

- They investigated within-session and cross-session developments of expertise, focusing on how the language and search behaviour of a user on a topic evolves over time.
- The paper demonstrates a connection between clicks and several metrics related to expertise.
- Based on models of the user and their specific context, the paper presents a method capable of automatically predicting, with good accuracy, which clicks will lead to enhanced learning.



Research Papers

Title 4. Language model adaptation using web documents obtained by utterance-based queries.[11]

Key Observations :

- They focused on the quality of the generated queries and propose a novel query generation method.
- In contrast to the n-gram based queries used in past works, their approach relies on utterances as candidate queries.



Research Papers

Title 5. POS Tagging of English-Hindi Code-Mixed Social Media Content. [12]

Key Observations :

- Code-mixing is frequently observed in user generated content on social media, especially from multilingual users. The linguistic complexity of such content is compounded by presence of spelling variations, transliteration and non-adherence of formal grammar.
- Their results show that language identification and transliteration for Hindi are two major challenges that impact POS tagging accuracy.



Research Papers

Title 6. Improving Performance Of English-Hindi Cross Language Information Retrieval Using Transliteration Of Query Terms. [13]

Key Observations :

- The main issue in Cross Language Information Retrieval (CLIR) is the poor performance of retrieval in terms of average precision when compared to monolingual retrieval performance.
- The main reasons behind poor performance of CLIR are mismatching of query terms, lexical ambiguity and un-translated query terms.
- They used all possible combination of Hindi translated query using transliteration of English query terms and choosing the best query among them for retrieval of documents.



Research Papers

Title 7. Shallow Parsing Pipeline for Hindi-English Code-Mixed Social Media Text. [14]

Key Observations :

- In this study, the problem of shallow parsing of Hindi-English code-mixed social media text (CSMT) has been addressed.
- They have annotated the data, developed a language identifier, a normalizer, a part-of-speech tagger and a shallow parser.
- **Dataset Example:**
 - hy... try fr sm gov job jiske forms niklte h...
 - **Gloss:** Hey... try for some government job which forms give out...
 - **Translation:** Hey... try for some government job which gives out forms...



Research Papers

Title 8. Identifying languages at the word level in code-mixed indian social media text. [15]

Key Observations :

- Language identification at the document level has been considered an almost solved problem in some application areas, but language detectors fail in the social media context due to phenomena such as utterance internal code-switching, lexical borrowings, and phonetic typing; all implying that language identification in social media has to be carried out at the word level.
- For word level language detection they have explored various aspects:
 - N-gram Language Profiling and Pruning
 - Dictionary-Based Detection
 - SVM-based Word-Language Detection
- To understand the effect of each feature and module, experiments were carried out at various levels. The n-gram pruning and dictionary modules were evaluated separately, and those features were used in the SVM classification.



Overall architecture

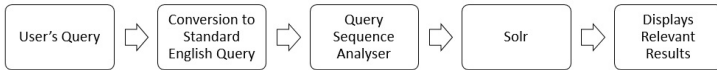


Figure: Schematic diagram of the pipeline



Module I : Conversion to Standard English Query



Figure: Hindi-English Code Mixed Query to Standard English Query



Relevance of this module

- Under the influence of social media, more and more people are using code-mixed query. For a bi-lingual speaker of Hindi-English, he is at ease typing the query as a mix of hindi and english terms.
- But the data is mostly available in English.
- The solution is two-fold:
 - Either convert the query to English.
 - Or convert the documents to be matched as per the query language.
- Are we doing Translation of the entire query? **NO!**



Dataset Examples I

- **Code-Switching** : Language switches from English to Hindi.
Eg: gov job jiske forms nikalte hain
- **Code-Mixing** : Some English words are embedded in a Hindi utterance.
Eg: divya bharti mandir marriage kendra kahan hai
- **Spelling variations** : sm - some, gov - government
- **Ambiguous Words** : To - so in Hindi or To in English



Dataset Examples II

Sl.	Example	Type
1	try fr sm gov job jiske forms niklte h...	Code-Switching from English to Hindi
2	divya bharti temple marriage center ko donate karna	Code-Mixing (English word is used in Hindi utterance.)
3	tum kab ja rahe ho bcoz thr is no tckt avalble	Code-Switching from Hindi to English
4	you to aab gone	Code-Mixing (Hindi words are used in English utterance.)
5	thoda toh overacting banta hai na	Code-Mixing (English word is used in Hindi utterance.)
6	tumne koi movie dekhi	Code-Mixing (English word is used in Hindi utterance.)
7	wht r u doing, mein free hun..	Code-Switching from English to Hindi

Table: Code-Mixed Dataset Example



Data Preparation

- Code-Mixed data was obtained from social media posts from the data shared in FIRE-2014 shared task on Transliterated search.
- The existing annotations on the FIRE dataset was removed, posts were broken down into sentences.
- 858 of those sentences were randomly selected.
- The Language of 63.33% of the tokens in code-mixed sentence is Hindi.



Data Distribution

language	Sentences
English	141 (16.43%)
Hindi	111 (12.94%)
Code-Mixed	606 (70.63 %)
Total	858

Table: Data distribution at sentence level.

language	All Sentences	Only CM Sentences
Hindi	6318 (57.05%)	5352 (63.34%)
English	3015 (27.22%)	1886 (22.32%)
Rest	1742 (15.73 %)	1212 (14.34%)
Total	11075	8450

Table: Data distribution at token level.



Language Identification

- Every word was given a tag out of three 'en', 'hi' and 'rest'.
- Words that a bilingual speaker could identify as belonging to either Hindi or English were marked as 'hi' or 'en'.
- The label 'rest' was given to symbols, emoticons, punctuation, named entities, acronyms, foreign words.
- We treated language identification as a three class ('hi', 'en', 'rest') classification problem:
 - The feature set comprised of **BNC** (normalized frequency of the word in British National Corpus), **LEXNORM** (binary feature indicating the presence of the word in the lexical normalization dataset), **HINDI_DICT** (binary feature indicating the presence of the word in a dictionary of 30,823 transliterated Hindi words), **NGRAM** (word n-grams).
- We modeled the language identification as a sequence labeling task, where we employed CRF into usage.



Result of Language Identifier Module

Features	Accuracy
BNC	61.26
+LEX_NORM	71.43
+HINDI_DICT	77.50
+NGRAM	93.18

Table: Feature Ablation for Language Identifier



Normalization I

- Once Language identification is done, we convert the noisy non-standard tokens (such as Hindi words inconsistently written in many ways using the Roman script) in the text into standard words.
- To fix this, a normalization module that performs a language specific transformations, yield the correct spelling for a given word is being built.
- Two language specific normalizers, one for Hindi and other for English/Rest, generate normalized candidates which were then ranked.
 - We obtained character alignments between noisy Hindi words in Roman script H_r to normalized Hindi words format H_n using GIZA++ on 30,823 Hindi words.
 - Next, a CRF classifier was trained over these alignments, enabling it to convert a character sequence from Roman to Devnagri using learnt letter transformations.



Normalization II

- Then the subnormalizer uses our developed *Spell Checker* and cross validated using SILPA libindic Spell Checker to compute the normalized word for a given input word.
- At last the normalized Hindi words were converted from Devnagri script to Correctly spelled English word using a standard **Hindi** → **English** Dictionary.
- A similar approach was used for English text normalization.
- Words with language tag 'rest' was left unprocessed.

Normalizer	Accuracy
Hindi Normalizer	78.25%
English Normalizer	69.98%
Overall	74.11%

Table: Normalizer Accuracy



Example

- Query : **hapy to see u here swagat hai !**
- Tokens : ['hapy', 'to', 'see', 'u', 'here', 'swagat', 'hai', '!']
- Language Tag : ['en', 'en', 'en', 'en', 'en', 'hi', 'hi', 're']
- English Normalizer : ['happy', 'to', 'see', 'you', 'here']
- Hindi Normalizer : ['welcome', 'is']



Future Work of Query Conversion Module

- The work till now converts individual noisy (unnormalized) code-mixed tokens to normalized english tokens.
- Work to assign them their POS tag. So that we can minimize the information loss while conversion to standard english tokens.
- Build a shallow parser for code-mixed English-Hindi search query.



Working of Query Sequence Analyzer

- Maintain a sliding window of size 10.
- Break the query into semantic units.
- Identify if there is a need of query reformulation. At present our system is able to handle the following types of queries:
 - Pronoun replacement: If the subsequent query have any pronoun then it is substituted with the noun from the previous query.
 - Name entity identification: If there is no pronoun in the subsequent query, but there is some named entity in the queries then it tries to combine the queries belonging to the same named entity.
 - Continuity of concept: The query is segmented into semantic phrases depending upon the mutual information score. Whenever the point wise mutual information between two consecutive words drop below a predefined threshold, a segment break is inserted.



Relevance of this Module

- Let the query sequence be:
 - Q. 1 iiit bhubaneswar
 - Q. 2 m.tech courses there
 - Q. 3 how to reach * bhubaneswar (* iiit missing)
- There is an inherent continuity in the query sequence!



Module II : Query Sequence analyser - Semantic Units I

- First, we need to figure out, if the query need to be refined?
- We start by segmenting each query into phrases. Query segmentation is the process of taking a user's search query and dividing the tokens into individual phrases or semantic units.
- We adopted the Mutual Information method (MI).
- A segmentation for a query is obtained by computing the pointwise mutual information score for each pair of consecutive words. More formally, for a query $x = x_1, x_2, \dots, x_n$

$$PMI(x_i, x_{i+1}) = \log \frac{p(x_i, x_{i+1})}{p(x_i)p(x_{i+1})}$$

- In the implementation, the probabilities for all words and n-grams are computed using the Microsoft Web N-Gram Service.



Module II : Query Sequence analyser - Semantic Units II

- A segment break is introduced whenever the point wise mutual information between two consecutive words drop below a certain threshold $\tau = 0.89$ (The threshold was so choosed to maximize the break accuracy).
- In addition to breaking the query into phrases, we also grouped multi-word keywords together (e.g. “new delhi”, “narendra modi”, etc.).
- We do that by adopting a hierarchical segmentation technique where the same segmentation method described above is re-applied to every resulting phrase with a new threshold $\tau_s < \tau$.
- We selected the new threshold ($\tau = 1.9$) to maximize the break accuracy over a set of a random sample of 10,000 Wikipedia title of persons, cities, countries and organisations.
- Now that we have phrases and keywords in each query, we assume that every phrase corresponds to a semantic unit.



Future Work in Query Sequence Analyzer Module

- Assigning the POS tag to the Query tokens.
- Identifying the head. It may be the last noun keyword.
- Identifying the keywords in a Query.
- Auto Query reformulation based on matching concepts.



Module III : Implementation in Solr

- After second stage refinement, we send the query to Solr.
- Solr is a popular open source search platform.
- For our project, we are indexing the pdf and doc files into our solr system. For this purpose we are using Apache Tika.
- Our solr system then returns ranked wise name of documents containing relevant information as per our query.
- To evaluate our system, we can calculate the precision and compare the results for original query and refined query.



More about Solr and Tika

- Apache Solr is an open source search platform built upon the Java library called Lucene.
- Solr is a popular search platform for Web sites because it can index and search multiple sites and return recommendations for related content based on the search query's taxonomy. Solr is also a popular search platform for enterprise search because it can be used to index and search documents and email attachments.
- Apache Tika is a content detection and analysis framework, written in Java.
- It detects and extracts metadata and text from over a thousand different file types.



Desktop Search Application

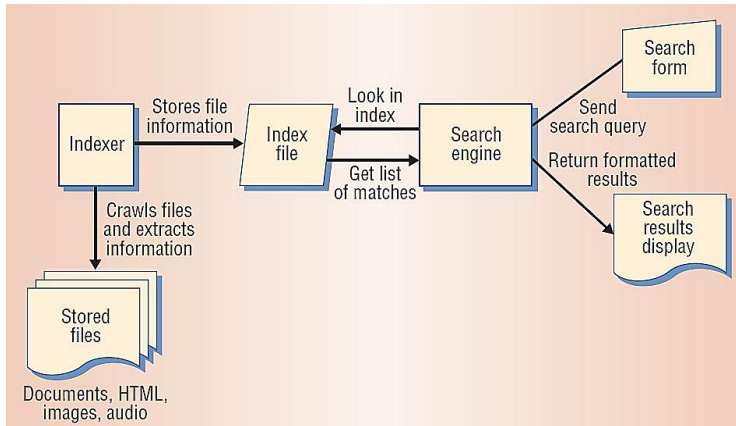


Figure: Schematic diagram of the Desktop Search Application



Desktop Search Application Results I

Sl. No.	Query	Path of Documents	P@5
Q1	machine learning aur jankari retrval	/media/amit/Academics/BOOKS /COMPUTERS/AI & ANN Books/book-neuro-intro.pdf	0.8
New Query	machine learn- ing and informa- tion retrieval	/media/amit/Academics/BOOKS /COMPUTERS/Information Retrieval/[Bing_Liu]_Web_Data _Mining_Exploring_Hyperlinks, _(BookFi).pdf	0.9
		/media/amit/Project/Machine Learning Books /181115.pdf	0.7
		/media/amit/Project/Machine Learning Books /RW.doc	0.75
		/media/amit/Project/Machine Learning Books/ ciml-v0_9-all.ppt	0.85

Figure: First Query



Desktop Search Application Results II

Sl. No.	Query	Path of Documents	P@5
Q2	usmein intellgnt systms	/media/amit/Project/Machine Learning Books/ IntroMLBook.pdf	1
New Query	machine intelli- gent systems	/media/amit/Project/Machine Learning Books/neuronalenetze-en-zeta2-2col-dkrieselcom.docx	0.9
		/media/amit/Project/Machine Learning Books/whole.pdf	0.8
		/media/amit/Project/Machine Learning Books/book.pdf	1
		/media/amit/Project/Machine Learning Books/book_draft.pdf	0.9

Figure: Second Query



Desktop Search Application Results III

Sl. No.	Query	Path of Documents	P@5
Q3	feature neekalna	/media/amit/Project/Machine Learning Books	0.8
New Query	feature extraction	/media/amit/Project/Machine Learning Books/book_draft.pdf	0.8
		/media/amit/Project/Machine Learning Books/ISLR Sixth Printing.pdf	1
		/media/amit/Project/IR-NLP/Semantic Search - Research Papers/rhodes-phd-JITIR.pdf	1
		/media/amit/Project/IR-NLP/Semantic Search - Research Papers/joachims_etal_05a.doc	0.7

Figure: Third Query



Precision against Google Results for Query I

The table below shows the precision values calculated for the results (Links) given by Google Search engine. To test our model against a large corpus, we manually entered the queries after reformulation in Google.

Query #	Original Query	P@3	P@5	Final Reformulated Query	P@3	P@5
Q ₁	bharat ki rajdhani	0	0	bharat's capital	0.6	0.5
Q ₂	capital of the country	0.5	0.35	capital of bharat	0.6	0.5
Q ₃	last viceroy of india	1	1	last viceroy of india	1	1
Q ₄	ind ka prim minister kaun hai	0	0.2	india's prime minister who is	1	0.95
Q ₅	unka kitna age	0	0	india's prime minister how much age	1	1
Q ₆	gov jobs jiska exam hota hai	0	0.2	india govt job whose exam have	0.65	0.7
Q ₇	festivals here	0	0	festivals india	1	1
Q ₈	mickel jackson fav step	0	0	mickel jackson favourite step	0	0
Q ₉	mumbai to kolkata kese jana hai	0.3	0.28	mumbai to kolkata how to go	0.6	0.73
Q ₁₀	what people here eat	0	0	what people india eat	0.6	0.5

Figure: Query sequence sent to our system




Precision against Google Results for Query II

Query #	Original Query	P@3	P@5	Final Reformulated Query	P@3	P@5
Q ₁₁	wannacry kya hai (Before updating the NER list)	0.3	0.2	want to cry what is (Before updating the NER list)	0	0
Q ₁₂	wannacry kya hai (After updating the NER list)	0.3	0.2	wannacry what is (After updating the NER list)	1	0.9
Q ₁₃	badrinath ke pas landslde	0.8	0.7	badrinath near landslide	0.9	1
Q ₁₄	wht hapnd thr	0	0	what happened badrinath	0	0.2
Q ₁₅	wahan ka temp kitna	0	0	Badrinath's temperature	0.8	0.7
Q ₁₆	zomato dat thft	1	1	zomato data theft	1	1
Q ₁₇	book fr mchine lerning	0.9	0.8	book for machine learning	1	0.9
Q ₁₈	places 2 vist at bbsr	0.7	0.6	places to visit at Bhubaneswar	1	0.8
Q ₁₉	wahn famous ky hai	0	0	bhubaneswar famous what is	0.8	0.8
Q ₂₀	Odisha assembly ka speaker	1	1	odisha assembly's speaker	1	1

Figure: Queries on Trending topics



Google Results for Original Query



[All](#) [Videos](#) [News](#) [Maps](#) [Images](#) [More](#) [Settings](#) [Tools](#)

About 1,45,000 results (0.36 seconds)

How to Prevent WannaCry - This Post Shows You How - tenable.com
[\(Ad\) www.tenable.com/tenable-io/container](#) ▼
Our solution architect Disney Cheng details the steps you need to take.
Reduce Exposure & Loss · Free Training · Prioritize Threats · Eliminate Blind Spots

दुनियाभर को डराने वाला साइबर अटैक क्या है और इससे ...
[hindi.firstpost.com](#) > [Technology](#) ▼ [Translate this page](#)
5 days ago - इस रैनसमवेयर वानाक्राइ (WannaCry) या वानाक्रिप्ट (WannaCrypt) का नाम ... क्या होता है रैनसमवेयर साइबर अटैक.


Wanna Cry Ransomware Cyber Attack Kya hai or Kaise Bache - Hindi ...
[https://hindimehelp.com/wanna-cry-ransomware-cyber-attack/](#) ▼
5 days ago - Abhi ek new Virus aaya hai jiska naam hai **Wanna Cry** Ransomware Virus jisse 2,00,000+ computers affected huee hai world wild, or ye abhi tak ka sabse bada cyber attack bataya ja raha hai. Chaliye Jaan lete hai akhir kya hai **Wanna Cry** Ransomware Cyber Attack or isse kaise bacha jaye ...

क्या है वानाक्राइ रैनसमवेयर वायरस - What is Wanna Cry ...
[www.mybigguide.com/.../What-is-Wanna-Cry-ransomware-Virus-...](#) ▼ [Translate this page](#)
12 मई 2017 को दुनियाभर में कंप्यूटरों पर एक वायरस का हमला हुआ है, जिससे भारत भी अछूता नहीं है इस वायरस का नाम है ...

रैनसमवेयर अटैक क्या है? - in 1 Hindi Lifestyle Blog for Everyone
[www.hindi.com/ransomware-virus-attack-and-prevention-tips-in-...](#) - [Translate this page](#)



Before Updating the NER List



[All](#)
[Images](#)
[Videos](#)
[News](#)
[Maps](#)
[More](#)
[Settings](#)
[Tools](#)

About 3,88,00,000 results (0.50 seconds)

How to Cry and Let It All Out: 14 Steps (with Pictures) - wikiHow
[www.wikihow.com > ... > Managing Sadness and Nostalgia > Crying](http://www.wikihow.com/Managing-Sadness-and-Nostalgia/Crying) ▼
 ★★★★★ Rating: 76% - 658 votes
 Jan 3, 2017 - Letting yourself feel emotions and letting yourself cry proves that you're a strong, independent woman because you aren't afraid of people seeing you actually have emotions. ... Remember, don't think about crying. Think about why you are sad. Don't do it to cry, do it because you need to deal with your emotions.


I feel so sad right now, I really want to cry. But in fact, I can't. So what ...
<https://www.quora.com/I-feel-so-sad-right-now-I-really-want-to-cry-But-in-fact-I-can...> ▼
 There's a DBT (Dialectical Behavioral Therapy) skill called "opposite action," in which you act exactly opposite to what your emotional state is telling you to do.

Why do I feel like crying but can't? - Quora
<https://www.quora.com/Why-do-I-feel-like-crying-but-cant> ▼
 We know what brings people to tears - sadness and pain, but also beauty and joy. Here it is in the words of a couple of psychologists quoted on WebMD: 'Crying is a natural emotional response to certain feelings, usually sadness and hurt.

The Inability to Cry | Psychology Today
<https://www.psychologytoday.com/blog/how-everyone-became.../the-inability-cry> ▼
 Apr 15, 2014 - But if you find that you can't cry, that you can't feel anything, what then? ... Browning captured this inability to cry In her 1844 poem "Grief," She ...



After Updating the NER List





[All](#)
[Maps](#)
[Videos](#)
[News](#)
[Images](#)
[More](#)
[Settings](#)
[Tools](#)

About 3,89,00,000 results (0.49 seconds)

Latest on WannaCry Ransomware - Keep Your Business Safe
[Ad](#) secure.f-secure.com/Ransomware/WannaCry ▼
 F-Secure tells you how to protect your business against Ransomware attacks.
[All Business Solutions](#) · [Subscribe to Our Blog](#) · [Cyber Security StressTest](#)

Wannacry (or WannaCrypt, WanaCrypt0r 2.0, Wanna Decryptor) is a ransomware computer worm that targets the Microsoft Windows operating system. The virus was used to launch the **WannaCry** ransomware attack on Friday, 12 May 2017.

WannaCry - Wikipedia
<https://en.wikipedia.org/wiki/WannaCry>


 About this result  Feedback

WannaCry - Wikipedia
<https://en.wikipedia.org/wiki/WannaCry> ▼
Wannacry (or WannaCrypt, WanaCrypt0r 2.0, Wanna Decryptor) is a ransomware computer worm that targets the Microsoft Windows operating system. The virus was used to launch the **WannaCry** ransomware attack on Friday, 12 May 2017.

What is WannaCry and how does ransomware work? - The Telegraph
www.telegraph.co.uk › [Technology](#) ▼
 2 days ago - The "WannaCry" ransomware appears to have used a flaw in Microsoft's software,



An Outlier...



[All](#)
[News](#)
[Videos](#)
[Images](#)
[Maps](#)
[More](#)
[Settings](#)
[Tools](#)

About 4,09,000 results (0.52 seconds)

A twist in the tale: Badrinath ki Dulhania review by Sarit Ray | movie ...
www.hindustantimes.com/...badrinath.../story-B02I8opByp2szSAoygVn5J.html ▼
 Mar 10, 2017 - **What happens** when the feisty girl stays feisty? Love meets feminism for a new kind of rom-com.

Badrinath Ki Dulhania - Wikipedia
https://en.wikipedia.org/wiki/Badrinath_Ki_Dulhania ▼
 Badrinath "Badri" Bansal (Varun Dhawan) is the younger son of a wealthy family in Jhansi. ... But when Badri helps Vaidehi's older sister, Kritika find a husband and even resolves a dowry crisis, Vaidehi agrees to marry him. On their wedding day, however, Vaidehi never shows. Badri is heartbroken and his father is furious.

Release date: 10 March 2017 (India) **Budget:** ₹35 crore
Starring: Varun Dhawan; Alia Bhatt **Hindi:** बद्रीनाथ की दुल्हनिया

Subramaniam Badrinath - Wikipedia
https://en.wikipedia.org/wiki/Subramaniam_Badrinath ▼
 Subramaniam **Badrinath** (born 30 August 1980) is an Indian cricketer. He is a right-handed middle order batsman. **Badrinath** has represented India in One Day International matches.

Test debut (cap 262): 6 February 2010 v South ... **Last Test:** 14 February 2010 v South Africa
ODI debut (cap 176): 20 August 2008 v Sri Lan... **Last ODI:** 13 June 2011 v West Indies



Dissemination Of Work

- 1 Amit Jena and Rakesh Chandra Balabantaray. "Query Optimization using Query Sequence for Hindi-English Code-Mixed Query." SCESM 2017.

Published in: International Journal of Control Theory and Applications.

- 2 Amit Jena and Rakesh Chandra Balabantaray. "Semantic Desktop Search Application for Hindi-English Code-Mixed user Query with Query Sequence analysis"

Presented in: the Conference IRTEICT 2017.






References I

-  Parth Gupta, Kalika Bali, Rafael E. Banchs, Monojit Choudhury, and Paolo Rosso. 2014. Query Expansion for Mixed-script Information Retrieval. In: The 37th Annual ACM SIGIR Conference, SIGIR-2014, Gold Coast, Australia, June 6-11, pp. 677-686.
-  Li, Xin, and Dan Roth. "Learning question classifiers." Proceedings of the 19th international conference on Computational linguistics-Volume 1. Association for Computational Linguistics, 2002.
-  Somnath Banerjee, Sudip Kumar Naskar, Paolo Rosso, and Sivaji Bandyopadhyay. 2016. The First Cross-Script Code-Mixed Question Answering Corpus. In: Modeling, Learning and Mining for Cross/Multilinguality Workshop, 38th European Conference on Information Retrieval (ECIR), pp.56-65.







References II

-  Royal Sequiera, Monojit Choudhury, Parth Gupta, Paolo Rosso, Shubham Kumar, Somnath Banerjee, Sudip Kumar Naskar, Sivaji Bandyopadhyay, Gokul Chittaranjan, Amitava Das, and Kunal Chakma. 2015. Overview of FIRE-2015 Shared Task on Mixed Script Information Retrieval. In: Forum for Information Retrieval Evaluation (FIRE), pp. 19-25.
-  Gamback, Bjorn, and Amitava Das. "Comparing the level of code-switching in corpora." Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). 2016.
-  Kunal Chakma, and Amitava Das. CMIR:A Corpus for Evaluation of Code Mixed Information Retrieval of Hindi-English Tweets. 2016. In: 17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING), April 3-9, Konya, Turkey.








References III

-  Anupam Jamatia, Bjorn Gambäck, and Amitava Das. 2015. Parts of $\hat{\Delta}$ Speech Tagging for Code English-Hindi Twitter and Facebook Chat Messages. In: 10th Recent Advances of Natural Language Processing (RANLP), September, pp. 239-248.
-  Brin, Sergey, and Lawrence Page. "Reprint of: The anatomy of a large-scale hypertextual web search engine." Computer networks 56.18 (2012): 3825-3833.
-  Lu, Zheng, et al. "A new algorithm for inferring user search goals with feedback sessions." IEEE transactions on knowledge and data engineering 25.3 (2013): 502-513.
-  Eickhoff, Carsten, et al. "Lessons from the journey: a query log analysis of within-session learning." Proceedings of the 7th ACM international conference on Web search and data mining. ACM, 2014.



References IV

-  Tsiartas, Andreas, Panayiotis Georgiou, and Shrikanth Narayanan. "Language model adaptation using www documents obtained by utterance-based queries." 2010 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2010.
-  Vyas, Yogarshi, et al. "POS Tagging of English-Hindi Code-Mixed Social Media Content." EMNLP. Vol. 14. 2014.
-  Varshney, Saurabh, and Jyoti Bajpai. "Improving Performance Of English-Hindi Cross Language Information Retrieval Using Transliteration Of Query Terms." preprint :1401.3510 (2014).
-  Sharma, Arnav, et al. "Shallow Parsing Pipeline for Hindi-English Code-Mixed Social Media Text." arXiv preprint arXiv:1604.03136 (2016).
-  Das, Amitava, and Bjorn Gamback. "Identifying languages at the word level in code-mixed indian social media text." (2014).

