Yioop! Introducing Autosuggest and Spell Check

Advisor/Committee –

Dr. Chris Pollett, Dr. Sami Khuri, Dr. Robert Chun

-Sandhya Vissapragada

Topics

- Introduction and preliminary work
- Basic autosuggestion in Yioop
- Enhancements to the autosuggest feature
- Autosuggest for foreign languages
- Suggestions using previous queries
- Spell correction for English
- Suggestions for transliterated queries

Introduction

- Autosuggestion provides a dropdown menu of choices below the textbox in which a user is typing
- Spell correction helps in correcting the wrongly typed query
- Popularly found in [2]
  - Web browsers – Suggests URLs
  - Search engines – Suggests relevant queries
  - Word processors – Suggestions are generally from a dynamic dictionary built using the words in the doc
  - Code editors – Helps in typing long programs, example, IDE Eclipse

- Aim was to add the autosuggestion and spell correction features to Yioop!
- They help in reducing the typing work and in correcting spelling errors
- Google Instant is a popular implementation
  - Runs machine clusters and uses lists of popular queries from their logs to provide relevant suggestions to users

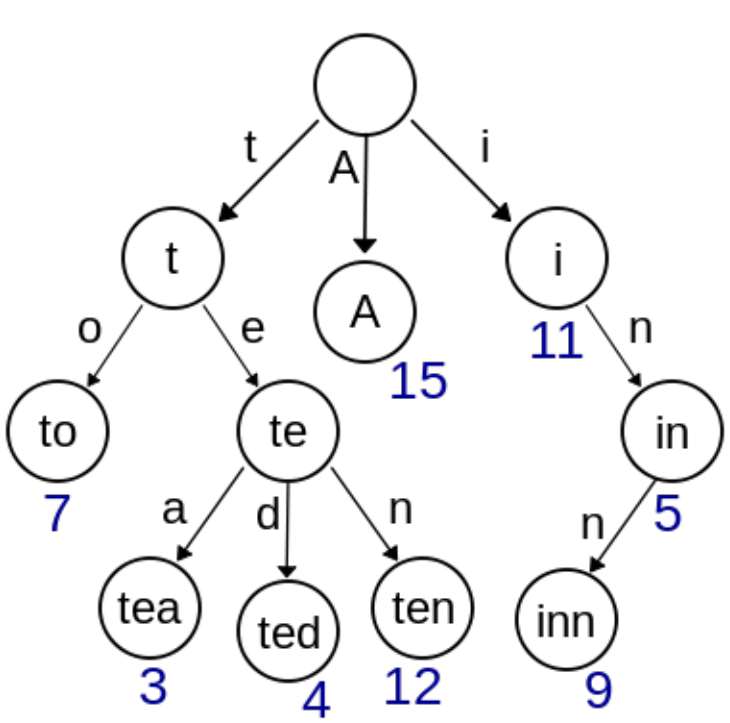
Yioop! & Constraints

- Yioop! - A PHP based search engine [1]
- Yioop runs on fewer machines
- Multiple server hits for these suggestions will reduce the performance
- There is no external user query data to rely on
- All the processing has to be done locally on the client machine

Storing dictionary words

- Comprehensive set of dictionary words have been chosen from wiki sources [5]
- Efficient storage of such huge data is crucial to avoid higher load times
- Trie is a suitable data structure
- Example of trie is shown in the next slide

Example of trie



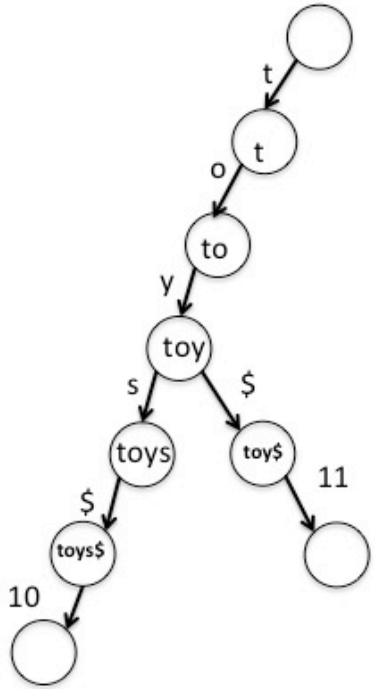


Figure 1 – Example of a trie

Initial steps to create a trie

- Trie was constructed using multi-level PHP arrays
- The trie was then JSON encoded and a gzip version was created.
- Words with less than 3 letters or stop words [8] or any words which has non-ASCII characters were discarded
- The final 250 KB gzip file was sent over the network and loaded when website was launched.

Timing tests

Firefox web console [9] was used

Trie type	Size in KB	Response time in ms
Plain JSON	*2500*	*2500*
Plain JSON with gzip enabled on HTTP	*2500*	*400*
File with compressed JSON data	*250*	*35*
Zipped JSON file with deflate option	*110*	*3*

Table 1 – Trie load time for different formats

Autosuggest in Yioop

- Initially word suggestions were incorporated in Yioop for US English
- Only English alphabet characters were handled
- It did the following –
  - *Trie was downloaded when the Yioop page was loaded.*
  - *On every, 'onKeyUp', a Javascript event, relevant suggestion words were retrieved and displayed.*
  - *Only the top six words are made visible*
  - *The user can hover the cursor on the suggestions and click one of them as the query.*
  - *Otherwise, the user can also use the arrow keys to traverse through the list and press the Enter key to submit the query*

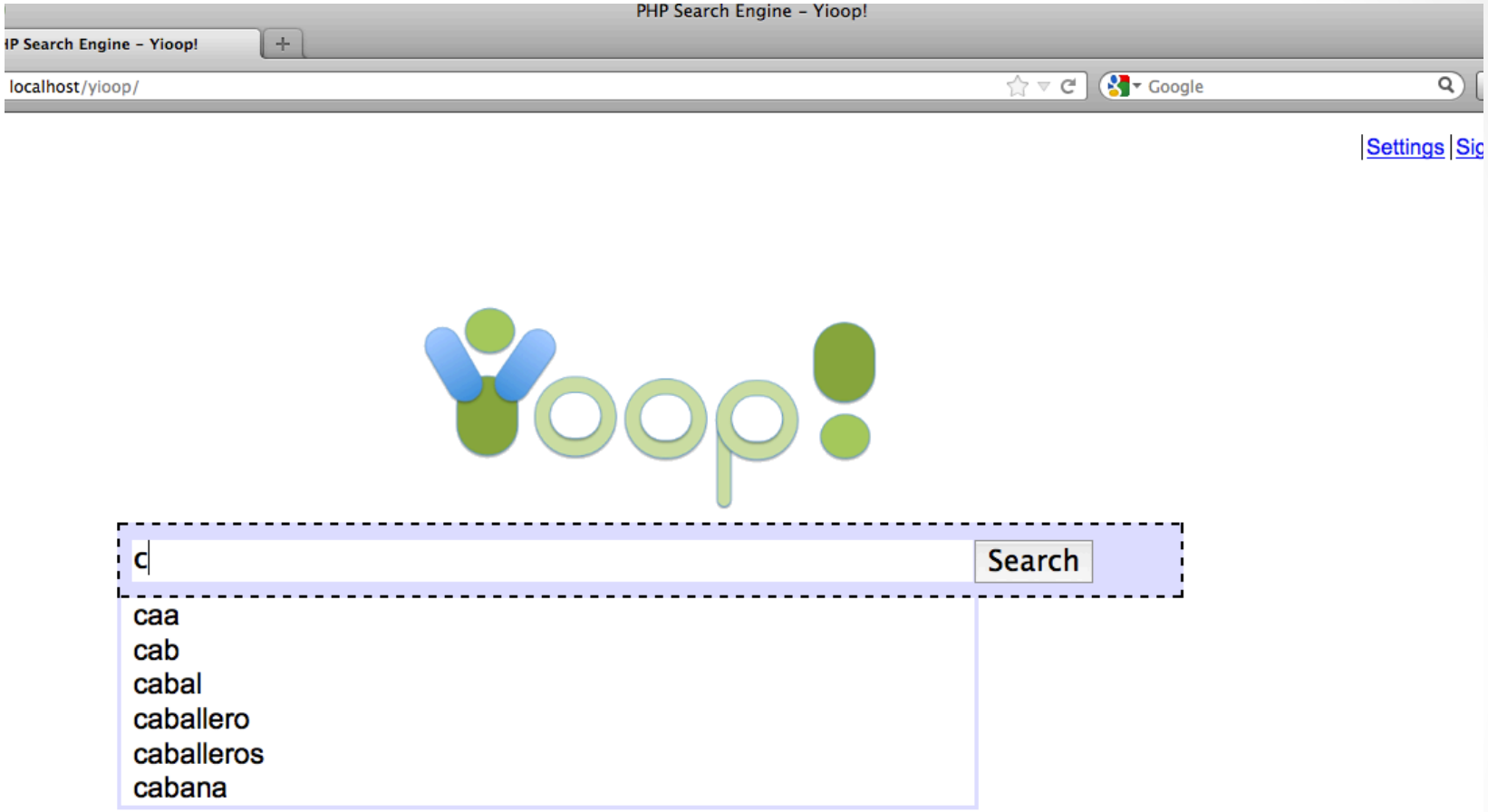


Figure 2 - Suggestions for character 'c'

Multi-word suggest

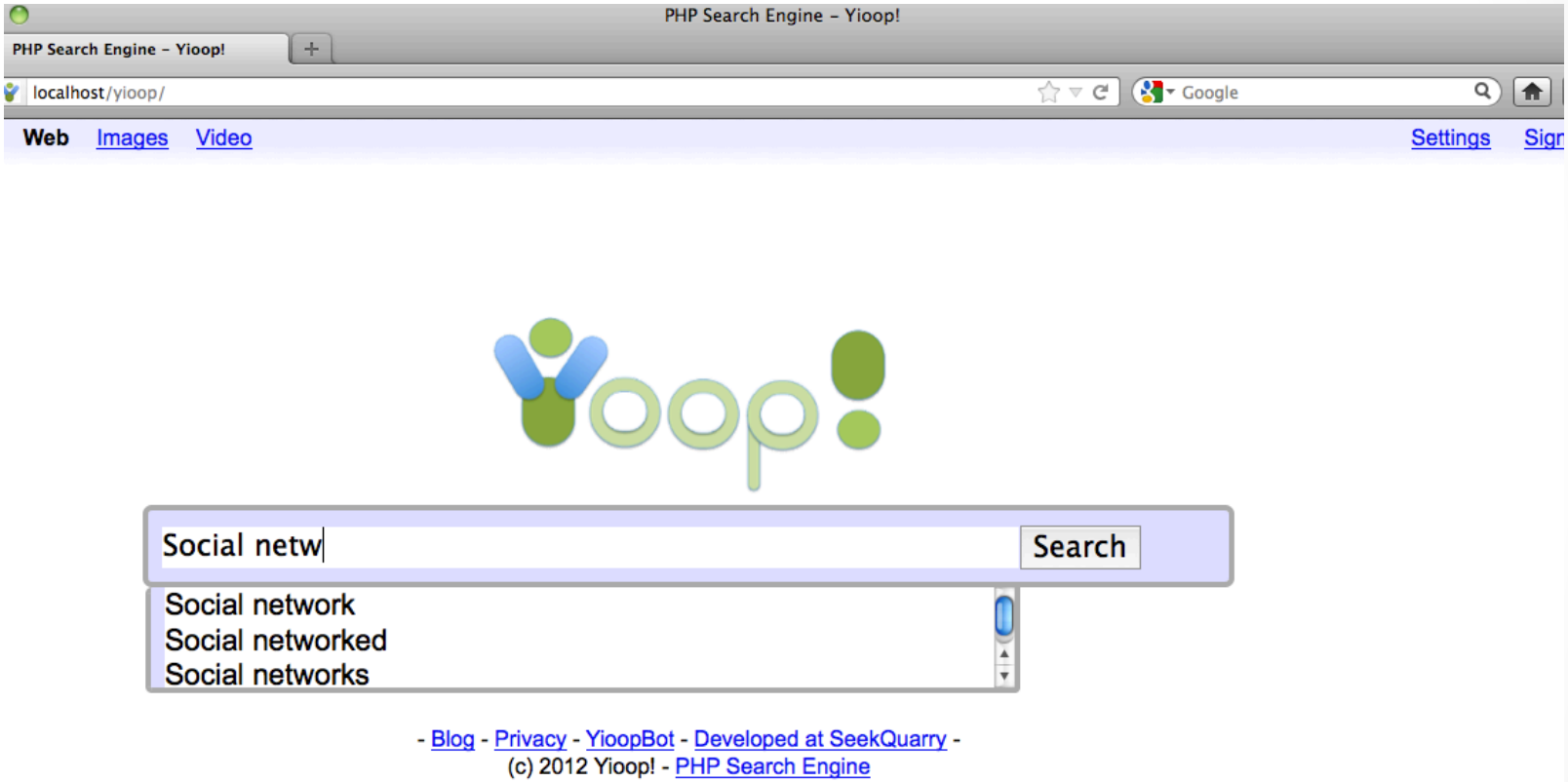


Figure 3 – Multi-word suggestions

Multi-word suggest

- Previous query terms are prepended to suggest a phrase
- Also, scroll bar was added to view more suggestions

Foreign language support

- With growing popularity of search engines, its important to support multiple languages
- Yioop! has a flexibility to add support to new languages
- Multiple byte data was handled to make it able to work for all sorts of inputs
- Now this feature supports any language with characters in the Unicode representation

Foreign language support

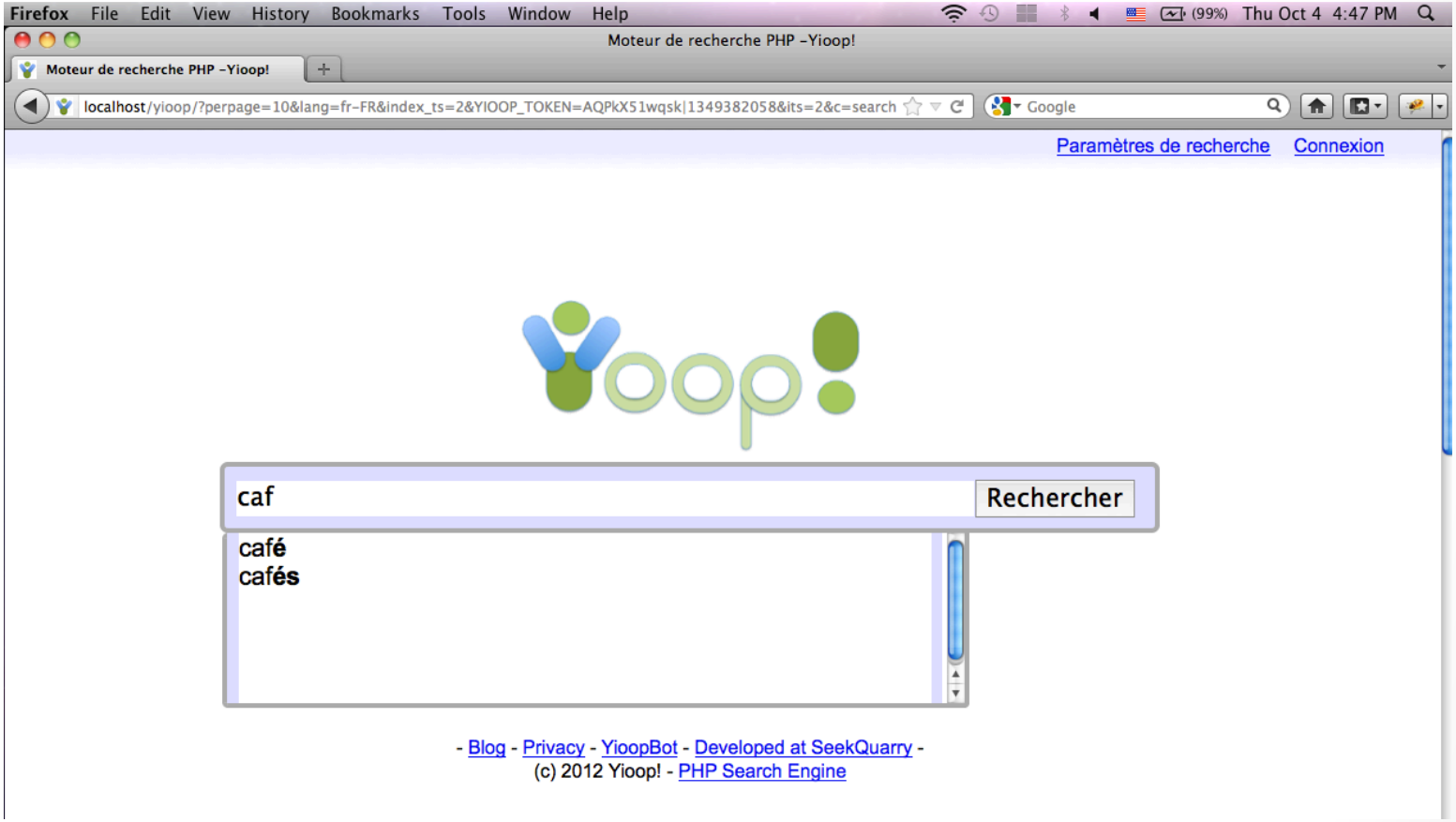


Figure 4 - Suggestions for French query

Suggestions using previous queries

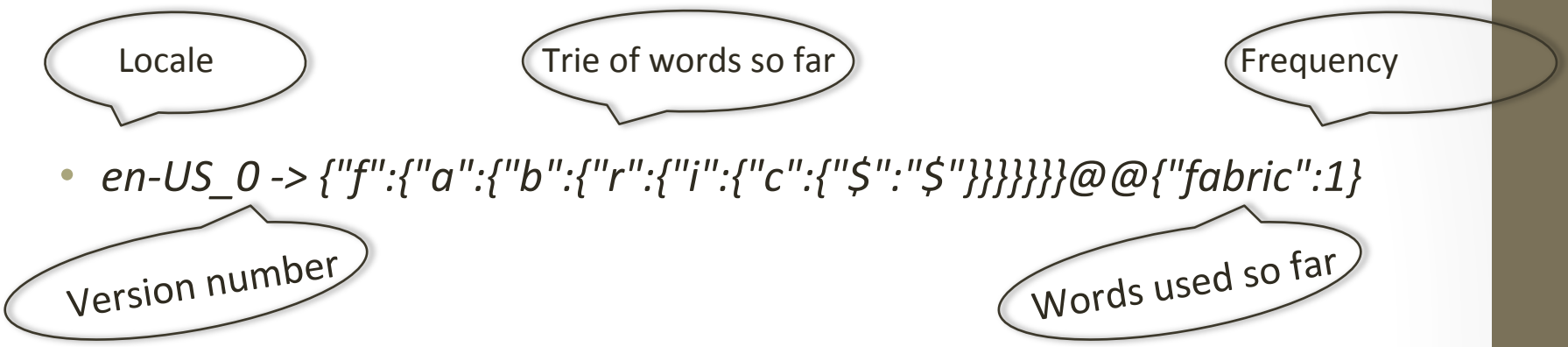
- Suggestions based on previous queries will be effective when same queries are typed multiple times

- Browser's local storage has been used in the form of key-value pairs [10]

Algorithm –

- Every query from a user along with its frequency (number of times the query was searched) is stored, specific to each 'locale'.
- Local storage words will top the suggestion list in the order of their frequency of occurrence

- The words are stored in the following fashion



Suggestions using previous queries

- When a user types a query the next time, first the local storage is checked for any existing suggestions
- If available, they appear first in the suggest list and are listed in descending order by the total number of times they have been fired.
- The actual dictionary is searched for further suggestions

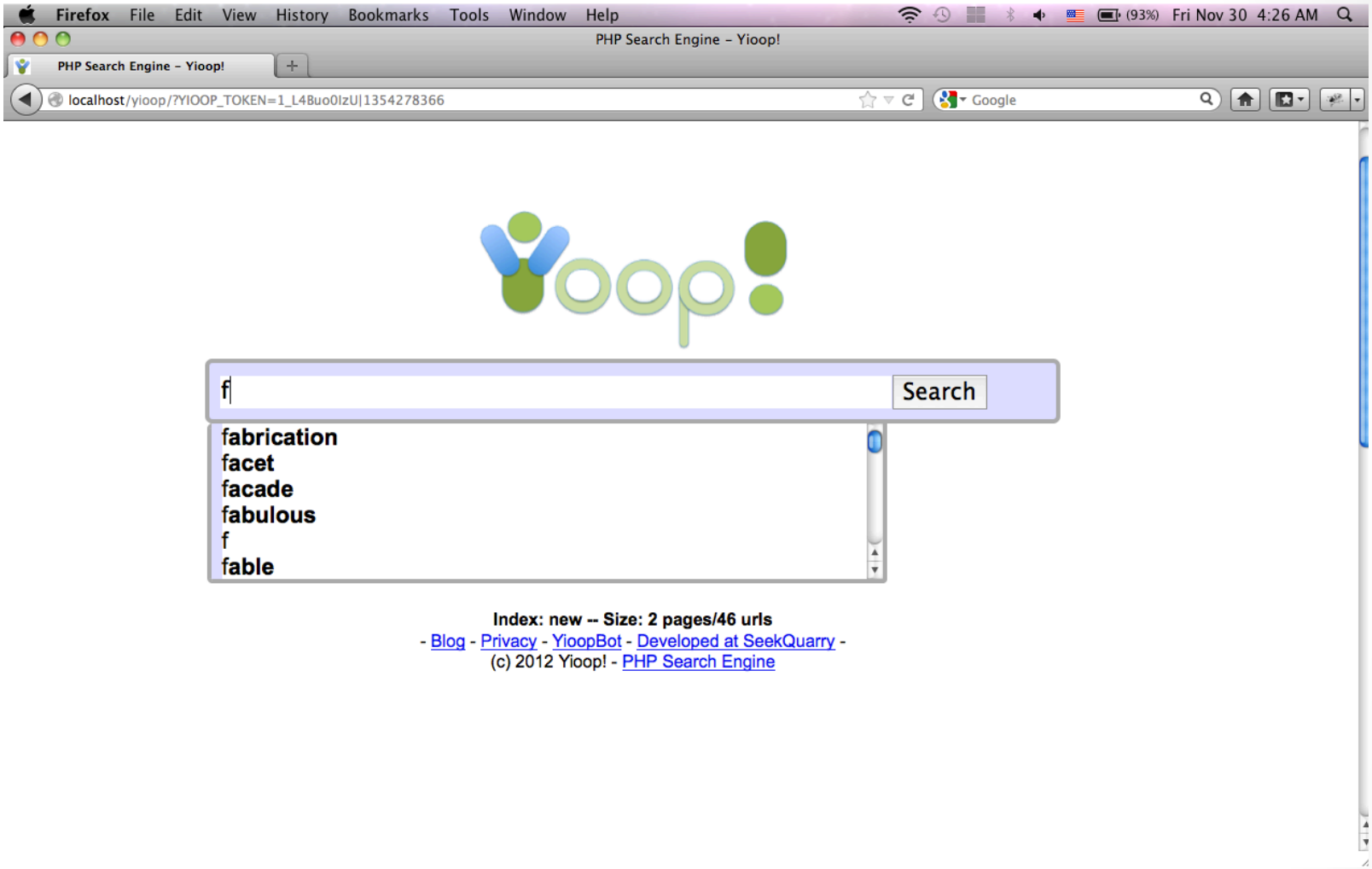


Figure 5 – Local storage example

Spell correction for English

- Helps user by correcting misspelled words, in turn saving time
- Google's 'Did you mean:' is a similar feature
- No external query data is available for Yioop, hence dictionary is used
- Dictionary structure was modified to have frequency of occurrence in the trie

Spell correction - Algorithm

- Edit distance algorithm is used [3]
- The number of edits it would take to turn into correct word is the edit distance between the two words.
- The possibilities are – [11]
  - A deletion where a letter is removed,
  - A transposition where there is a swap of adjacent letters,
  - A replacement where another replaces a letter or
  - An insertion where an unwanted letter is inserted

Spell correction - Algorithm

SpellCorrection (word):

Candidates = known (word) or known (Edits1 (word)) or word

Return candidate with maximum frequency in the trie

Edits1 (word):

Deletes: Set of words with one letter deleted

Transposes: Set of words with a swap between the adjacent characters

Replaces: Set of words with every letter replaced by 25 other letters in English alphabet

Inserts: Addition of an unwanted letter at all given positions in a word

Known (words):

Returns the set of words that are in the dictionary

Trade-off

- About 90% of spelling errors are of edit distance 1, as claimed by the literature on spelling correction [11]
- But it is also quite possible that the spelling errors are an edit distance of 2.
- For the word ‘improve’ the candidates with edit distance – 1 will be is 390.
- The number of candidates when edit distance 2 is applied to the word ‘improve’ is 162,150.

Query	*Corrections*	
	Edit Distance 1	*Edit Distance 2*
tha	*the*	*the*
lagh	*laugh*	*last*
sceince	*science*	*since*
nees	*news*	*been*
latre	*later*	*are*

Table 2 – Comparison of Edit Distance 1 & 2

- With one letter errors, edit distance 1 gave better results.
- To avoid the computational overhead of edit distance 2 algorithm, only edit distance 1 is chosen

Suggestions for transliterated queries

- Transliteration is the process of mapping text written in one language in to another by means of a pre-defined mapping [13].
- Its common to use English transliteration for foreign languages.
- Users who do not know that script of a particular language, tend to use this method.
- In the case of unavailability of a direct method to input data in a given language, transliteration becomes handy

Telugu – English transliteration

- Telugu is the third most spoken language in India, has been chosen. Telugu has 56 letters (18 vowels and 38 consonants) [4]

- Every phoneme in Telugu script when transliterated using English, ends with a vowel.

- Based on this, the approach of constructing a mapping table has been chosen.

telugu_array['k']='క';

telugu_array['kh']='ఖ';

telugu_array['+aa']='ా';

telugu_array['+oo']='ూ'

Telugu – English transliteration

- Assumption – The query typed in English should be a widely accepted transliteration
- The input query is divided into chunks based on the criteria of end character being a vowel
- Eg – manasu is divided as Ma + Na + Su
- These are then mapped against the mapping table to generate a Telugu query which is further processed for suggestions

Figure 6 – Suggestions for transliterated Telugu query

Performance

- Experiment 1: Queries were typed in the Yioop search box in following two modes.
    - ‘Word Suggest’ option disabled
    - ‘Word Suggest’ option enabled
- Following are the times recorded for five different people with different typing speeds.

Experiment 1

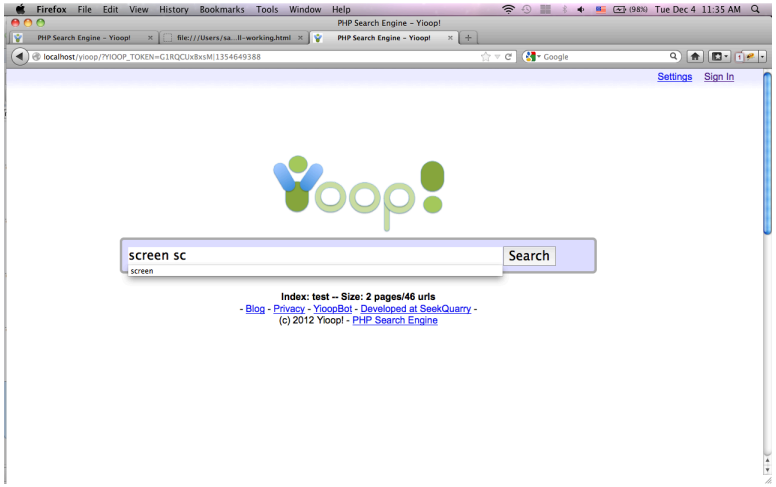
Word	Without Autosuggest – Time in sec					With Autosuggest – Time in sec				
	Per 1	Per 2	Per 3	Per 4	Per 5	Per 1	Per 2	Per 3	Per 4	Per 5
Science	4	5	4	4.5	3	2	3	2	3.5	2
Computer	5	4	4.5	4.5	3	3	2.5	3	3	2
Adapt	3.4	3.5	3	3	2	3	3	2	2	1.5
Accomplish	5	4	4.5	4.5	3	3	2.5	3	3	2

Table 3 –Experiment to test performance of autosuggest

Experiment 2

- To compare autosuggest against browser's autocomplete feature :
- Even this experiment involves two modes as mentioned above.
- Word suggest is turned off and the queries 'screen', 'scare' and 'science' have been searched 5 , 3 and 3 times respectively.
- Later, the same searches have been done in Yioop with the 'Word Suggest' option on. The following are the outcomes.

Experiment 2



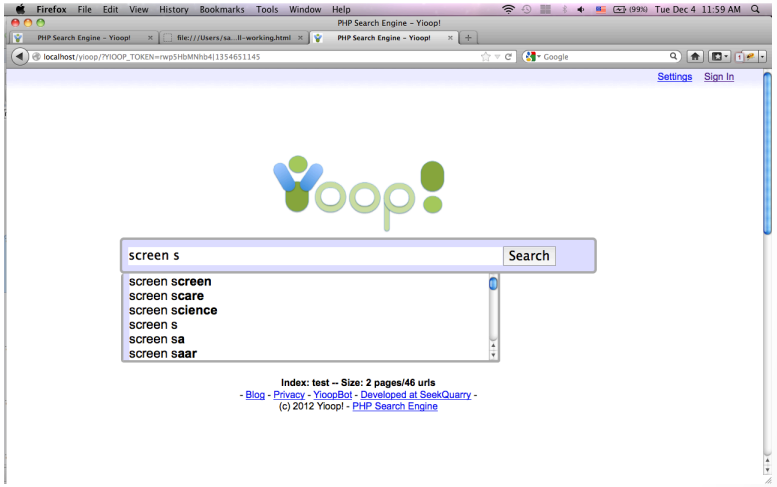
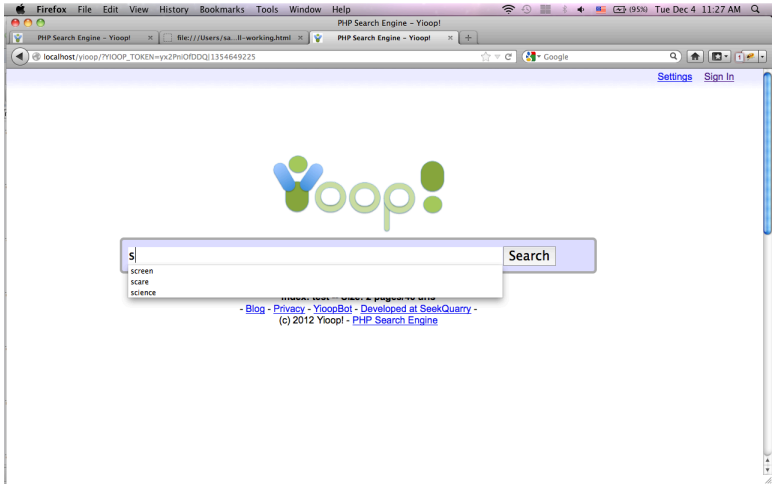


Figure 7 – Comparison of multi-word suggest with browser autocomplete

Experiment 2



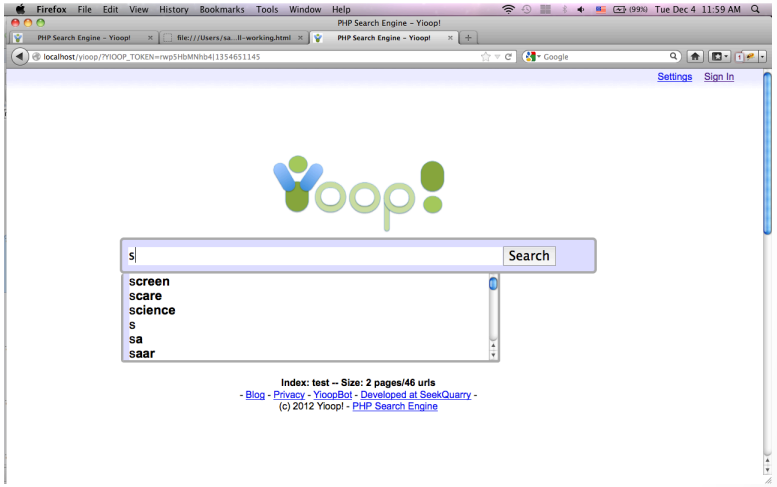


Figure 8 – Comparison of local storage suggest with browser autocomplete

Experiment 2

Foreign language support:

- Foreign language queries are supported by both
- Yioop additionally suggests words from dictionary

Spell correction:

- There is no spell correction in the browser
- Yioop saves retyping work as the search is just a click away

Summary

- Autosuggest and spell correction have been implemented for Yioop, keeping in mind its constrained environment
- Implemented in Javascript, autosuggest includes features like multi-word suggest, foreign language support and usage of locally stored previous queries, to enhance the performance
- Spell correction was implemented for English language, assuming that the frequency of one-letter errors is more than multiple-letter errors
- An attempt was made to suggest queries for English transliterated Telugu queries
- These features proved to reduce the typing work and correct spelling errors in Yioop.

Future work

- Using search result data for better suggestions as its more likely that an index is reused
- Introducing spelling correction for foreign languages like, French and Russian
- Introducing suggestions for transliterated queries pertaining to languages other than Telugu

References

[1] Yioop website: www.yioop.com Retrieved Nov 30, 2012

[2] Autosuggest: http://en.wikipedia.org/wiki/Autocomplete Retrieved Nov 30, 2012

[3] Edit distance: http://en.wikipedia.org/wiki/Levenshtein_distance Retrieved Nov 30, 2012

[4] Telugu: http://en.wikipedia.org/wiki/Telugu_language Retrieved Nov 30, 2012

[5] Yioop documentation: http://www.seekquarry.com/?c=main&p=documentation Retrieved Dec 4, 2012

[6] Popular English words: http://books.google.com/ngrams/datasets Retrieved May 15, 2012

[7] Trie: http://en.wikipedia.org/wiki/Trie Retrieved May 15, 2012

[8] Stop words: http://en.wikipedia.org/wiki/Stop_words Retrieved Dec 3, 2012

[9] Firefox web console: https://developer.mozilla.org/en-US/docs/Tools/Web_Console Retrieved Nov 30, 2012

[10] Local storage: http://www.w3schools.com/html/html5_webstorage.asp Retrieved Nov 30, 2012

[11] Spell correction: http://norvig.com/spell-correct.html Retrieved Nov 30, 2012

[12] Telugu writing system: http://en.wikipedia.org/wiki/Telugu_language#Writing_system Retrieved Nov 30, 2012

[13] Transliteration based Text Input Methods For Telugu, V.B. Sowmya and Vasudeva Varma , *22and International Conference on Computer Processing for Oriental Languages" 2009.*

[14] Telugu dictionary: http://www.lib.uchicago.edu/e/su/southasia/to_vijay/gwynn.txt Retrieved Nov 30, 2012

Thank you!