

Implicit User Profiling while Traversing Web

Introduction

Problem Statment

- User of search engines encounters a problem of getting vain data related to their searches.

for example: "Jaguar"

Jaguar + automobile OR Jaguar + mammal of felidae family

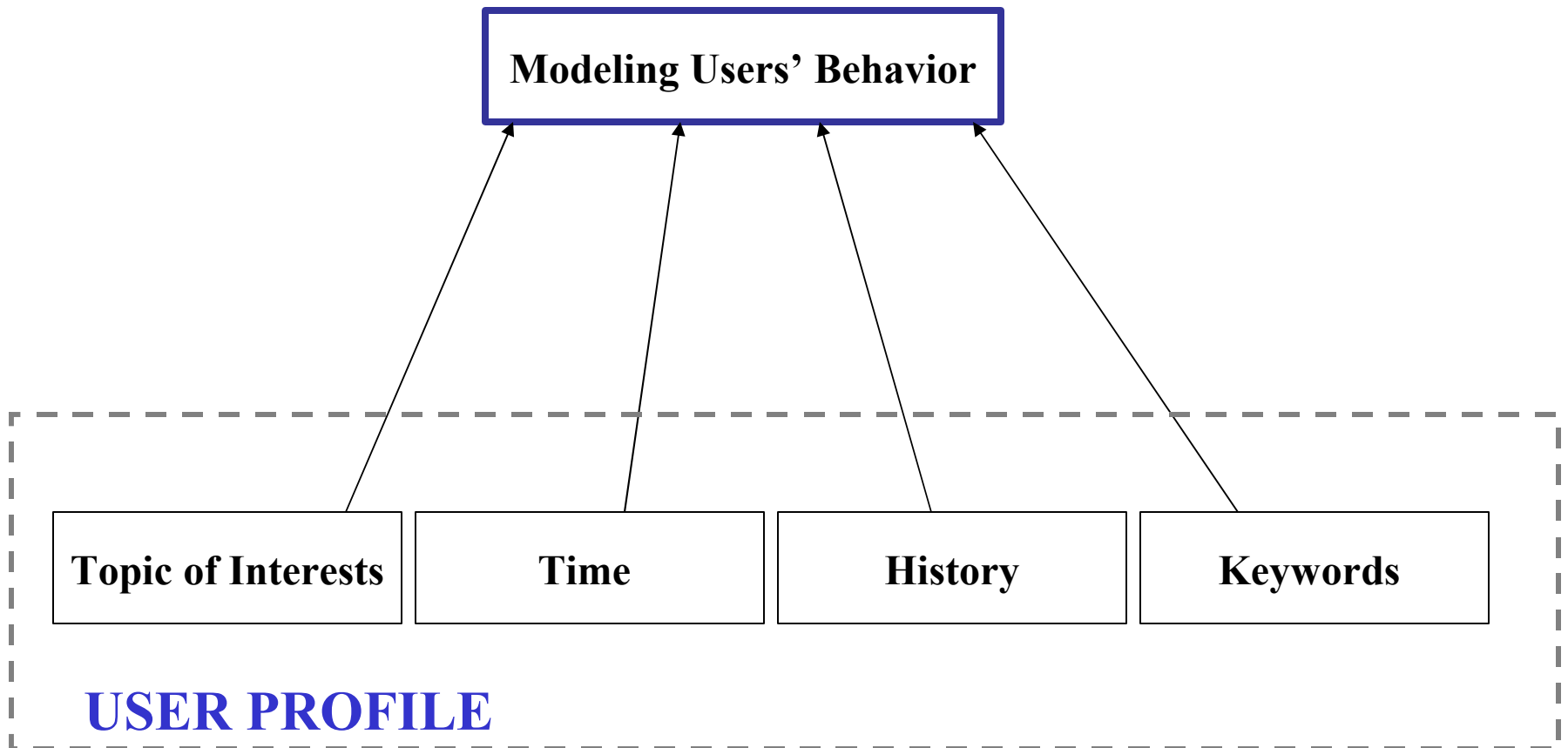
- Search engine usually displays some irrelevant data, which requires lot of time and effort from the user to parse information of their interest.
- So, need to develop a system that can act as:

SEARCH ENGINE:

Context + terms = OUTPUT TERM

Thesis Objectives

Building a context aware system requires

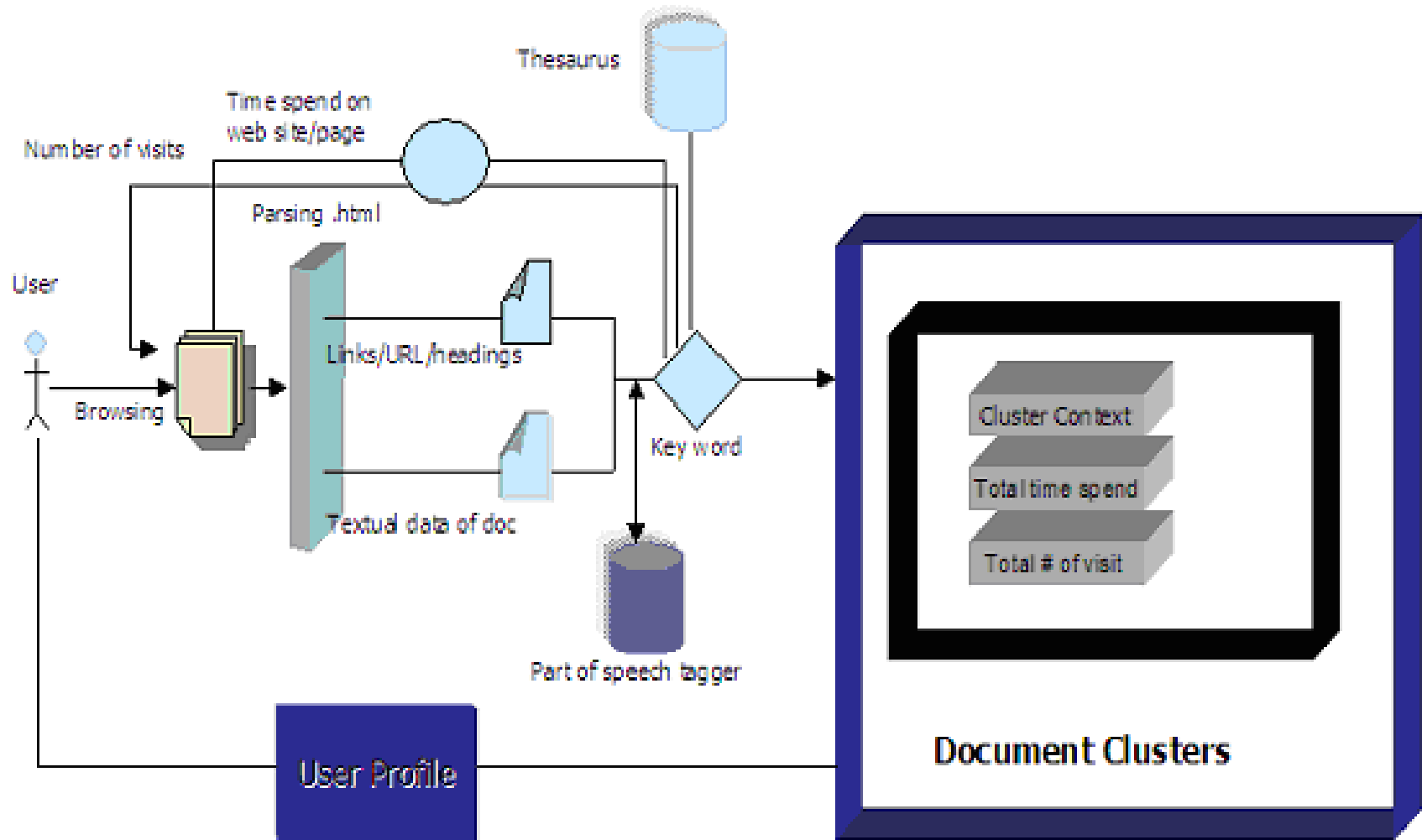


System Objectives

To build a user profile in web-based environment that will monitor user behavior and model it by,

- enabling to browse the web page.
- storing time spend by the user on a particular page.
- parsing html page and extract textual information.
- storing number of visit on the same page.
- retrieving keywords from the text.
- clustering webpage according to its context.

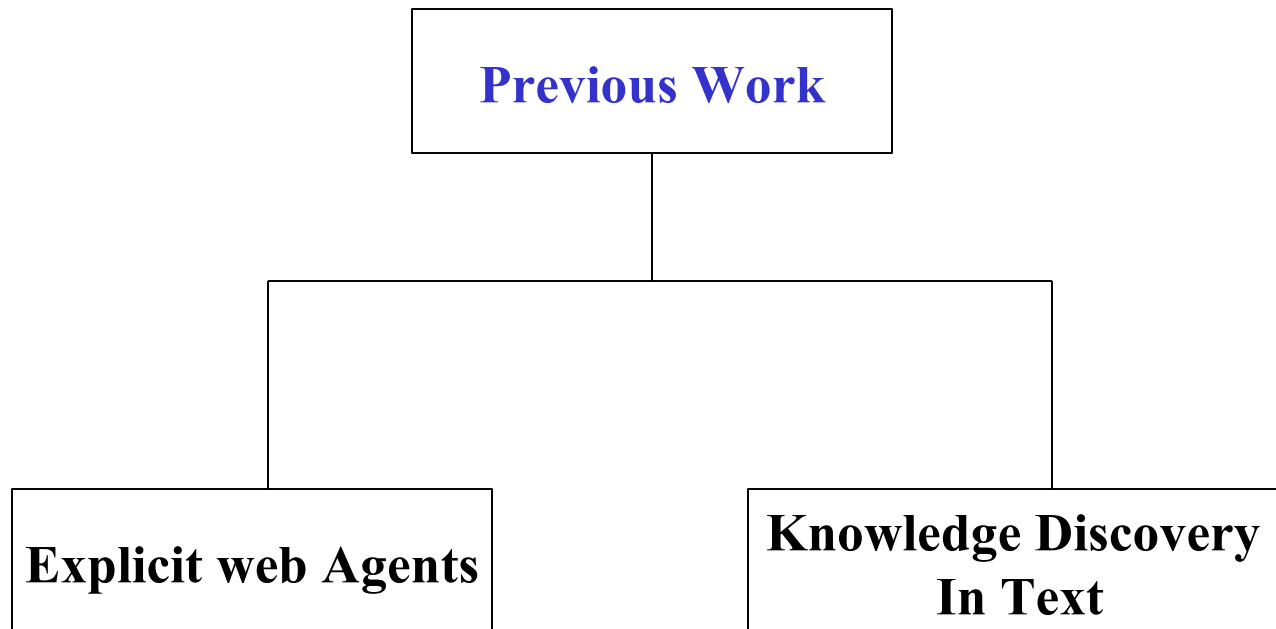
Proposed System Overview



Previous Work

Little about Previous Work

- Most of the work that we have found can be divided into two different forms,



System Details

Introduction to External Components

- HTML Parser
- Negative Dictionary
- Part of Speech Tagger
- WordNet Thesaurus

Phase I:

Knowledge Discovery in Text

HTML Parsing

- The system let the user browse web page, the web page is then send to html parser.
- Html parser parse the documents and output the data in the form of set of links , headings and paragraphs.

Stopping and Stemming

- After parsing the document, the textual data is sent to the process of stopping and stemming.

For Stopping:

Extendable Dictionary

For Stemming:

Lovins' Stemmer Algorithm

The suffixes dictionary we use is of max length 11 and provided by lovin.

Extracting Key Words

- Solving Word Sense Disambiguation
- Part of Speech Tagger
- Selecting Bigrams
- Thesaurus

Word Sense Disambiguation

- The nature of textual data is skewed and unstructured.
- There are words that have different semantics in different context

for example capital can be,

capital → capital city of a country

capital → capital investment by company

Part of Speech Tagger

- QTAG part of speech tagger is used.
- It will produce all information related to the grammatical structure of words.
- We restrict ourselves to

Noun

Verb

Adjectives

Bigrams

- Bigrams selection are based on the set of rules, these rules are simple and contains only noun, adjectives and verbs.
- These set of rules are made by combining two syntactic structures of words

Noun Verb

Noun Adjective

Adjective Noun

Verb Noun

Selecting Bigrams

- To calculate the association and strength of bigram we use Dice Coefficient

$$D_{ij} = \frac{2 \times x_{ij}}{Z + X}$$

- Shannon Diversity is used get the marginal distribution of a pair through the range of paris.

$$X = t_{nj} \cdot \log_{nj} - \sum_{i=1}^x t_{nji} \log t_{nji}$$

$$Y = t_{mi} \cdot \log_{mi} - \sum_{j=1}^x t_{mij} \log t_{mij}$$

Phase II:

User Profiling and Clustering

Clustering

- For clustering documents we decided to rank the keywords and for that we consider Inverse Document Frequency.

$$I_{df} = A_{b(occure)} \log \frac{D}{\sum_{i=1}^n D_{ib}}$$

- To find the similarity between different document Vector space model has been used.
- For clustering documents we use K- mean clustering algorithm.

User Profiling

For developing users' profile we considered following parameters.

- **Time Value** : total time spend by the user on web pages that belongs to certain cluster.
- **Cluster Size** : Number of documents, contain by a cluster.
- **Visit Score** : Frequency of visit on particular document.
- **Time on each web page** : Time socre related to one document.

Systems' Output Structure

System output

<i>Output structure of System</i>	
<i>Cluster_{number}</i>	$\left(Size = \sum_{numofpages=1}^n \right) \quad \left(\sum_{totalvisit}^n = T \right)$
1. <i>web document address</i>	$\sum_t^n Total\ time\ Spend \quad V_{count} = \sum_{count=1}^n$
2. <i>web document address</i>	$\sum_t^n Total\ time\ Spend \quad V_{count} = \sum_{count=1}^n$
3. ⋮ N..	

Conclusion
&
Future Work

Questions ??