TOWARDS
SEMANTIC WEB

by

SK. Saad Bin Kader
Exam Roll: Curzon Hall–527
Registration No: H–897
4^{th} year(Honors)
Session: 2008–2009

A Project submitted in partial fulfilment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
UNIVERSITY OF DHAKA

July 2021

Declaration

I, hereby, declare that the work presented in this project is the outcome of the investigation performed by me under the supervision of Sajib Kumar Mistry, Lecturer, Department of Computer Science and engineering, University of Dhaka. I also declare that no part of this project has been or is being submitted elsewhere for the award of any degree or diploma.

Countersigned

....

(Sajib Kumar Mistry)

Supervisor

Signature

....

(SK. Saad Bin Kader)

Abstract

The World Wide Web is a huge source of information and knowledge. But most of these are of little use when it comes to automated manipulation of the data because of their unstructured nature. Nowadays approach towards storing web content semantically is providing a great deal of meaningful data which can be used for automated data manipulation. Wikipedia, social networks like Twitter, Facebook, Google knowledge directory are some of the huge semantic web data. We are having a rapid increase of web users here in Bangladesh but unfortunately, the local Bangla websites are very unstructured. Even some of the best search engines are inefficient for retrieving data from these web contents. So a semantic approach towards building structured web content from now on will provide opportunities to the automated machines to manipulate more meaningful data. Guiding users towards more meaningful queries will certainly enhance the efficiency of search engines. Our goal is to create semantic knowledge on Bangla keywords to guide users for more efficient web data retrieval.

Acknowledgements

I consider myself very lucky to be a part of the Department of Computer Science and Engineering, the University of Dhaka where I have got the facilities to work on my project.

In the first place, I express my gratitude to my supervisor Sajib Kumar Mistry for allowing me to work under his inspiring supervision. I have learned a lot from him, not only from a scientific point of view but also a lot about eCommerce and social points of view. With him, I have had many extremely interesting and stimulating discussions. He encouraged and helped me in everything I did for my project, with his patience, understanding, and caring. It has been most pleasant to work with him.

I would like to thank my friends for giving me unconditional support. Without their support, I could have failed to finish my project on time. I am also very grateful to my office authority for giving me space for my project work.

Finally, very special thanks go to my family. In particular, I thank my father and my mother for giving me all the supports I ever needed.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

The World Wide Web is a huge collection of data that has a tremendous amount of data on thousands of categories. But most of these valuable data are stored in an unstructured way and mostly documented with human native language. So it is difficult for automated systems to retrieve, process, and integrate these data meaningfully. The traditional approach of data discovery relying on natural language processing is too much expensive and complex which varies a great deal from language to language. So we need a more efficient approach rather than depending on the ability to detect the grammatical pattern to retrieve meaningful data from text. The most common modern approach to solve these problems is to let the author represent the content more meaningfully. The approach is generally called web semantics. A semantic web provides more opportunities to the automated systems to have more meaningful data. Alongside web data mining, semantic representation enables powerful structured web content storage. Rather than wasting efforts on detecting natural language structures, the automated systems now can emphasize processing the available semantic contents rapidly and effectively. This proactive approach is saving a lot of time and effort providing more value to the contents.

Some of the biggest collections of web data(Wikipedia, Facebook, Twitter, Google, etc.) are constructed and maintained semantically which enables these huge collections efficiently accessible to millions of users all around the world. For the last few years availability of the internet and awareness of information technology resulting in the rapid growth of internet users here in Bangladesh. Language support for Bengali provided revolution to web content-based knowledge construction in Bengali. Informative website building, Blogging, adding content to Wikipedia and social networking in Bengali now became the latest trend. But most of these large collections of data are unstructured. Bringing these large growing collections of data under semantics will surely enrich them for future automation like translating to another language or language-independent information schema.

1.2 Motivation

Compared to Bengali, other languages like English have more structured web content and better language support for retrieving meaningful knowledge from the internet. Features like Google's knowledge-based web content search do not have to support Bengali. So searching Bengali content in the search engines lacks a great deal compared to other languages which have language support. This opens a lot of fields for thesis work and project developments to provide Bengali with these language supports.

1.3 Problem Definition

Users looking for Bengali content via the World Wide Web do not get the language support for retrieving desired results. Most of the provided results by the search engines contain too much irrelevant content because search engines have little or no knowledge of the keywords that are being searched. The complex nature of Bengali grammar making it very difficult and costly to parse data from the web contents by machine.

1.4 Overview of the Book

The book has five chapters titled Introduction, Background and Field Study, Proposer Approach, Features of Proposed Model, and Conclusion.

Chapter 1 gives an overview of the current problems which are faced by the users while searching for Bengali content. It also gives a specific view on the problem which we are trying to solve.

Chapter 2 highlights the study on the background of the problem and some related field studies. It also highlights the study on Bengali websites.

Chapter 3 discusses the proposed model to minimize the problem.

Chapter 4 provides a review of the features of developed web applications and its user interfaces.

Chapter 5 gives an overview of the tools which are used to develop the project.

Chapter 6 has the summary of the research project and fields where improvements are possible in the future.

Chapter 2

Background and Field Study

2.1 Unstructured Web Data

The World Wide Web is the biggest collection of information that is ever documented by man. The main reason for the rapid growth is its decentralized nature. Different hosts all over the world have different representations of their web content. Mostly because this huge collection was not designed to be processed by machines. As the size grew larger, finding the desired content became equivalently difficult. The simplicity and availability of web interfaces attracted more and more people towards it. Most of them have little knowledge of information technology. So they end up being documented information such a way that is hard for the automated systems to extract meaningful information out of them.

Nowadays, the use of the web is not limited to just storing and retrieving information. The users now have more expectations from it. They expect web applications to process complex queries for them and provide a more meaningful representation of the information stored on the web. However, the information stored on the web pages is only for representing the content in a human-understandable way. They provide very little information for the machines to parse meaningful data out of it. This leaves a huge gap between the expectation and what the users get.

Some of the automated systems provide support for retrieving meaningful information from natural languages but they are too much language-specific. Each natural language has its unique structure. Even the same language has different structures depending on the geographic location. So analyzing these complex structures of natural language and designing a language-specific automated system is too much expensive.

2.1.1 Search Engines and Social Networks

Search engines depend on crawling the web pages from the web directories and matching the contents with the query string. So, when we search keywords like "DU", the search engine does not know whether it is "Dhaka University" or Unix command "du" or even something else. This results in a lot of irrelevant results or no results at all. Also, the results are sorted mostly on the hit count of the websites which can lead to desired data below so much unwanted stuff. Web contents are found in so many different languages nowadays. Search engines are unable to produce language-independent results.

Social networks are a huge source of casual user information. Users have a link to each other but mostly their produced data are of no use. Wikipedia, blogs have a certain level of ontology but that is too abstract. Only a small percentage of the data has metadata to be able to get processed by a machine. Also, the representation of these data is too language-specific. So, without contribution from the author, it's too much unrealistic to process these data by automated systems.

2.1.2 Bengali Websites

Language support and writing tools like "Avro" prompted many web contents to be represented in Bengali. The number of internet users in Bangladesh is growing rapidly due to the development of information technology and the availability of the internet. People now can access the internet even from their smartphones. Among them, a large number of users use the internet for blogging and social networking in Bengali. So, there is a huge opportunity of building very large web knowledge on the Bengali language. Introducing semantic in Bengali website is relatively easier because it is still in its early days on the World Wide Web.

Because of the unstructured nature of Bengali websites, users face many difficulties. It is very hard for the users to find the right information from these sites. Even some of the most popular search engines fail to provide the users with relevant data. Lack of efficient web directories for Bengali websites, no support for natural language processing and information discovery for Bengali language, insufficient or no knowledge on information technology of the users resulting in a very bad web surfing experience for the users who are looking for specific information. Also, users of other languages have no way to retrieve data from Bengali websites which violates the main principle of the World Wide Web.

Being a relatively new language, Bengali websites have a really low hit count. So on the current ranking algorithm of the popular search engines, it is very hard to have these sites at the top of the list even they contain reasonable information about the topic. The blogging content management systems provide some level of ontology on the content of blog posts but for being a tree structure and very narrow relation between the parent and child it is very hard to discover a relationship among the contents. For this reason, automated systems fail to extract meaningful data from these contents.

Chapter 3

Proposed Approach

3.1 Introduction

The wide gap between the natural language (human language) and formal language (language understanding by computers) makes it impossible to extract meaningful data from most of the web contents. But the growing size of the World Wide Web is demanding automated systems to not only find information from the internet but also filter, refine and translate them into more meaningful information. Languages like English which are the main medium of representing most of the web contents on the internet have well-defined and relatively simple grammatical structures. English has the most similarity with the high-level computer languages. For these reasons designing automated systems to parse meaningful knowledge from the contents written in English is relatively easier. However, it is the opposite for languages like Bengali. Compared to English Bengali has a far more complex structure. Also considering the total users on the internet Bengali users are very few in number. So there is no step taken to construct automated knowledge generation for Bengali earlier our approach. Bengali web contents are relatively less in number and for their complex structure, they are not suitable for automatic parsing to extract meaningful information. For this reason, constructing knowledge on the keywords reactively rather than proactively is a better choice.

3.2 Model

Knowledge graphs which are represented by the mean of labeled nodes and labeled links between the links can be viewed as semantic networks. However one of the essential differences between knowledge graph and semantic network is, in knowledge graph various information about the nodes are not represented at all. This approach is accepted because of the following reasons-

1. Knowledge graphs are seen as models. Building a model is done by gradually extending the model which was simple at first.

2. Even a single type leads to specific considerations. So large number of relation types used in semantic networks introduce extra complication.

3. A small number of relation types allow only simple paths among the nodes. New relationships are constructed by inference. Many types of links can complicate the paths.

3.2.1 Ontology

Ontology is the backbone of semantic web structure. The concept of ontology is to formally represent knowledge as a set of concepts with in a domain. It provides relation between pair of concepts which can be used by the machine as meta data of those concepts. Model of the ontology could be in memory, in file or in database depending on the system that is designed with the ontology.

3.2.2 Knowledge Graph

Knowledge graphs are collection of tokens and types having well defined relationships among them.

1. **Token:** A token is a node in a knowledge graph. Everything those we experience in the real world or existent concepts in our mind can be expressed by tokens.

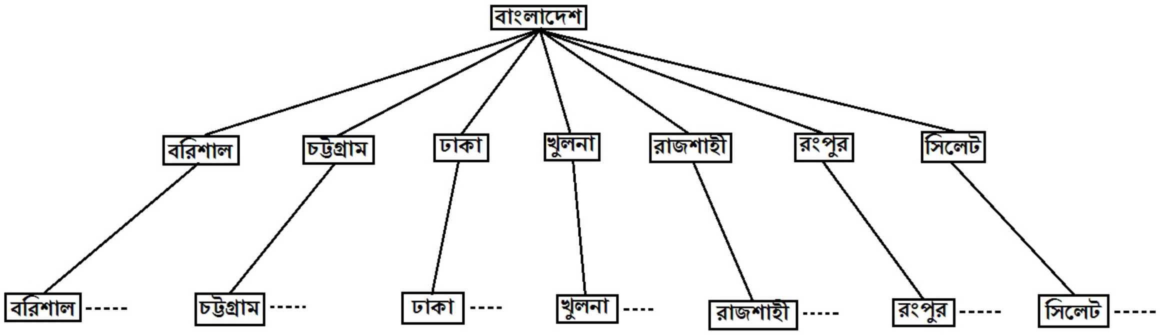


FIGURE 3.1: Ontology of Country, Division and District Class

2. **Types:** The basic nature of perception is to divide tokens into similar classes. There are identical tokens in the same class which can be expressed by introducing types for these tokens.

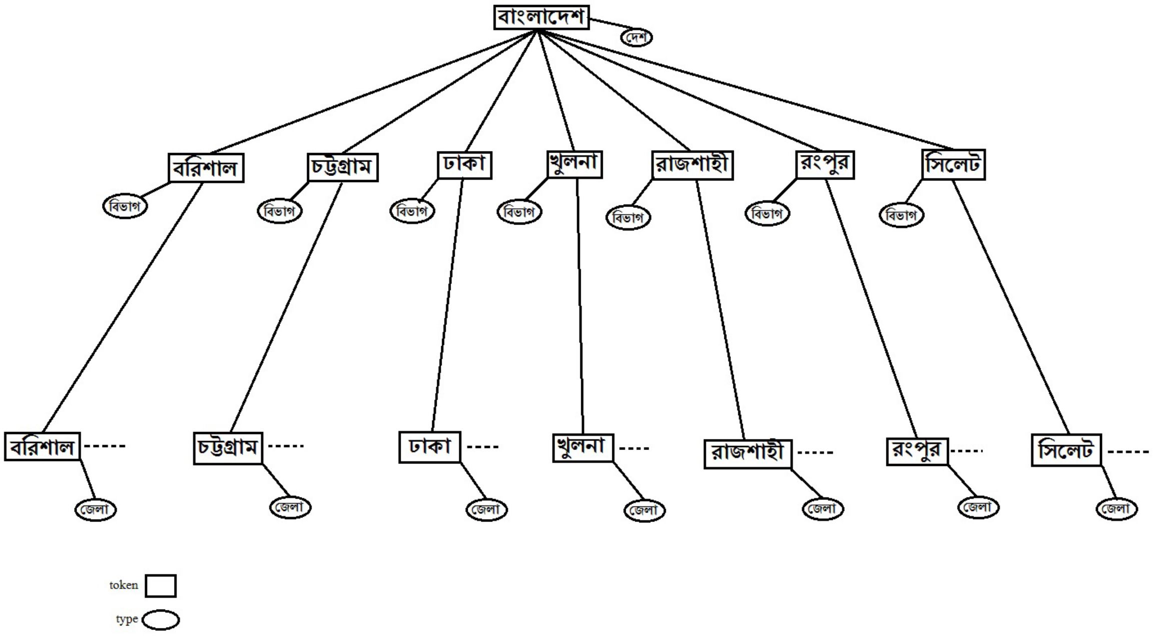


FIGURE 3.2: Knowledge Graph constructed on Country, Division and District Class

3.3 Guiding User Search

When users make a query they enter related key word for retrieving results. Then the search engine looks for pages those contain the key word and rank according to the hit count of the pages. This approach produces many irrelevant results because the search engine does not know the meaning of the key word or has very few information to determine the meaning of it. Our approach is to use knowledge graph to provide the search engine some extra meaningful information to filter its search result. This is done by looking in the knowledge to find information about the key word and pass it with the search key to the search engine. This produces significantly better result on many cases. The total procedure in done according to the algorithm on figure . . . .

3.4 Learning form User Search Evaluation

The initial state of the knowledge graph stats with fairly small size and having few token-type pair. But the main method of expanding our knowledge graph is via evaluating user search pattern. There can be cases when some user cannot find any information on the key word on the knowledge graph but the user will eventually find the desired result on their own effort filtering through the results provided by the search engine. We construct knowledge to the knowledge graph evaluating this information discovery process. So when next time some users have the same key word as query they will find information on the knowledge graph. The learning process of the knowledge graph depending on user search evaluation plan is done according to the algorithm on figure . . . .

3.5 Summary

Chapter 4

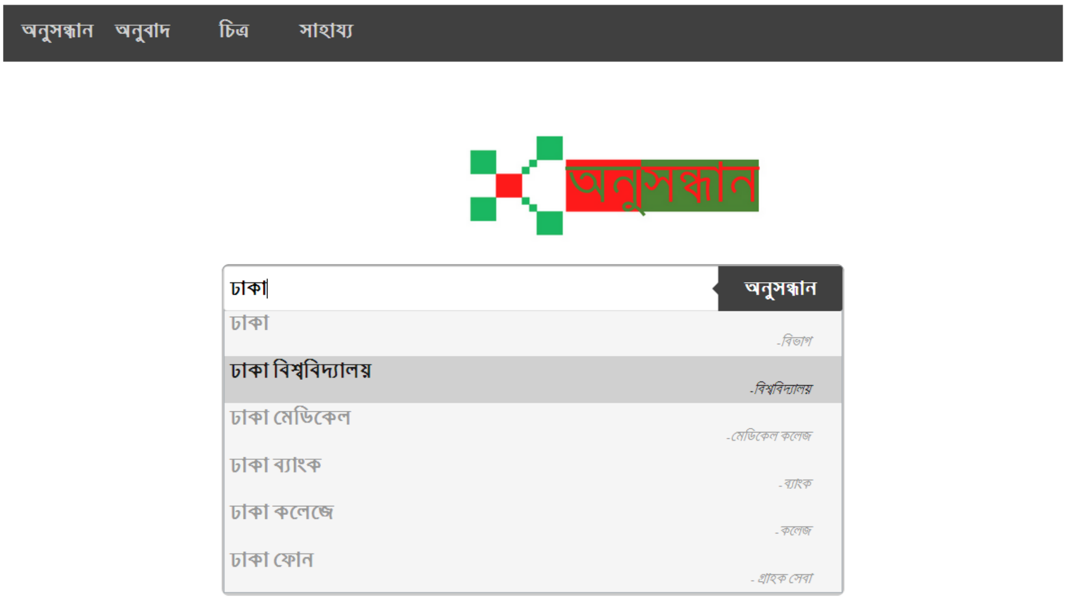
Features of Developed Model

4.1 Search Guidance

The main feature of our developed model is to provide users with guidance with the help of knowledge graph. With every keystroke suggestion results are generated from the knowledge graph. Generated suggestions get filtered as the size of the search key grows. When users click on a suggestion the key word get automatically adjusted according to the knowledge of that token-type pair.

4.2 Web and Image Searching

The application provides both web site and image results based on the key word that is searched. If the user adopts one of the suggested information, the search result is constructed on the basis of that. Otherwise default result construction of the search engine is shown. Both the web and image search results provide a title, a link to the original content and some part of the content from the page.



ঢাকা

ঢাকা -বিভাগ

ঢাকা বিশ্ববিদ্যালয় -বিশ্ববিদ্যালয়

ঢাকা মেডিকেল -মেডিকেল কলেজ

ঢাকা ব্যাংক -ব্যাংক

ঢাকা কলেজে -কলেজ

ঢাকা ফোন - গ্রাহক সেবা

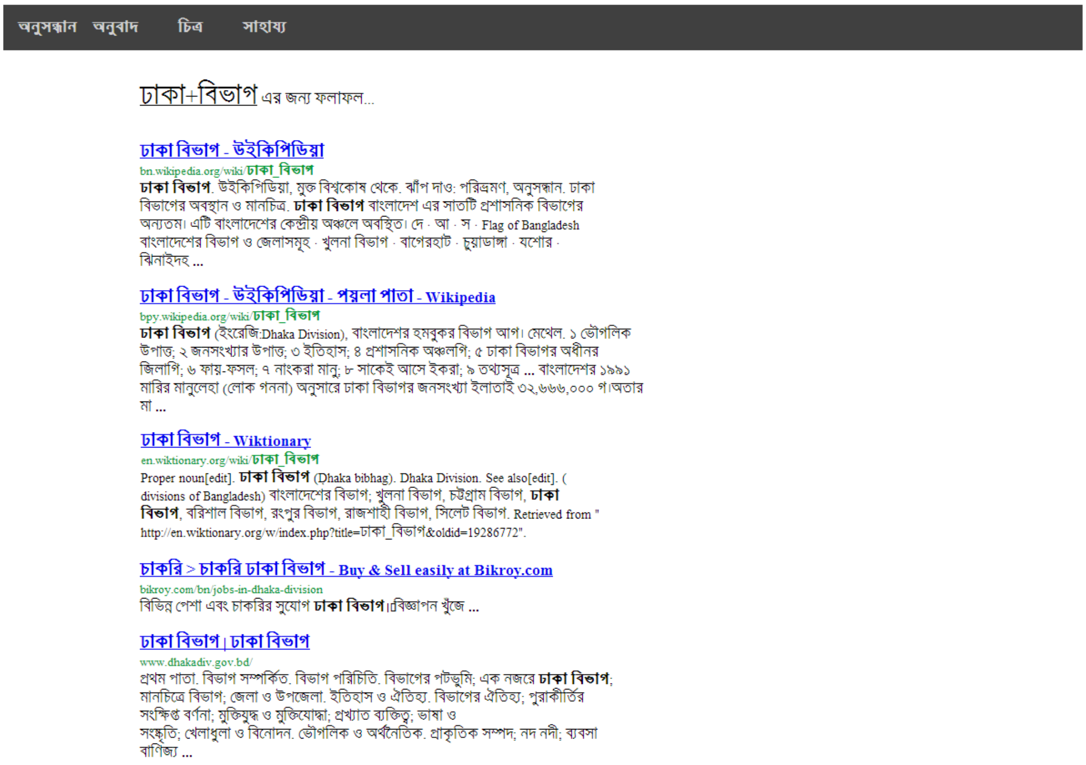
FIGURE 4.1: Search Suggestions

4.3 Translation

The users can use this option to translate an English word to Bengali. An online English to Bengali dictionary is used for translating the words. Users can also use the translated words for guided searching.

4.4 Expending Knowledge Graph

Users contribute to the growth of the knowledge graph. When the knowledge graph has exact knowledge about the search key, it provides the user with the information. But when it does not have information, it learns from the information provided by the users. The application can only learn about a token-type pair if the users provide them in a certain way. Here the '+' operator is used to represent the relation between a token-type pair. The system can learn about only one token-type pair at a time.



ঢাকা+বিভাগ এর জন্য ফলাফল...

ঢাকা বিভাগ - উইকিপিডিয়া
bn.wikipedia.org/wiki/**ঢাকা_বিভাগ**
ঢাকা বিভাগ. উইকিপিডিয়া, মুক্ত বিশ্বকোষ থেকে. ঝাঁপ দাও: পরিভ্রমণ, অনুসন্ধান. ঢাকা বিভাগের অবস্থান ও মানচিত্র. **ঢাকা বিভাগ** বাংলাদেশ এর সাতটি প্রশাসনিক বিভাগের অন্যতম। এটি বাংলাদেশের কেন্দ্রীয় অঞ্চলে অবস্থিত। দে · আ · স · Flag of Bangladesh বাংলাদেশের বিভাগ ও জেলাসমূহ · খুলনা বিভাগ · বাগেরহাট · চুয়াডাঙ্গা · যশোর · ঝিনাইদহ **...**

ঢাকা বিভাগ - উইকিপিডিয়া - পয়লা পাতা - Wikipedia
bpy.wikipedia.org/wiki/**ঢাকা_বিভাগ**
ঢাকা বিভাগ (ইংরেজি:Dhaka Division), বাংলাদেশর হমবুকর বিভাগ আগ। মেথেল. ১ ভৌগলিক উপাত্ত; ২ জনসংখ্যার উপাত্ত; ৩ ইতিহাস; ৪ প্রশাসনিক অঞ্চলগি; ৫ ঢাকা বিভাগর অধীনর জিলাগি; ৬ ফায়-ফসল; ৭ নাংকরা মানু; ৮ সাকেই আসে ইকরা; ৯ তথ্যসূত্র **...** বাংলাদেশর ১৯৯১ মারির মানুলেহা (লোক গননা) অনুসারে ঢাকা বিভাগর জনসংখ্যা ইলাতাই ৩২,৬৬৬,০০০ গ।অতার মা **...**

ঢাকা বিভাগ - Wiktionary
en.wiktionary.org/wiki/**ঢাকা_বিভাগ**
Proper noun[edit]. **ঢাকা বিভাগ** (Dhaka bibhag). Dhaka Division. See also[edit]. (divisions of Bangladesh) বাংলাদেশের বিভাগ; খুলনা বিভাগ, চট্টগ্রাম বিভাগ, **ঢাকা বিভাগ**, বরিশাল বিভাগ, রংপুর বিভাগ, রাজশাহী বিভাগ, সিলেট বিভাগ. Retrieved from " http://en.wiktionary.org/w/index.php?title=ঢাকা_বিভাগ&oldid=19286772".

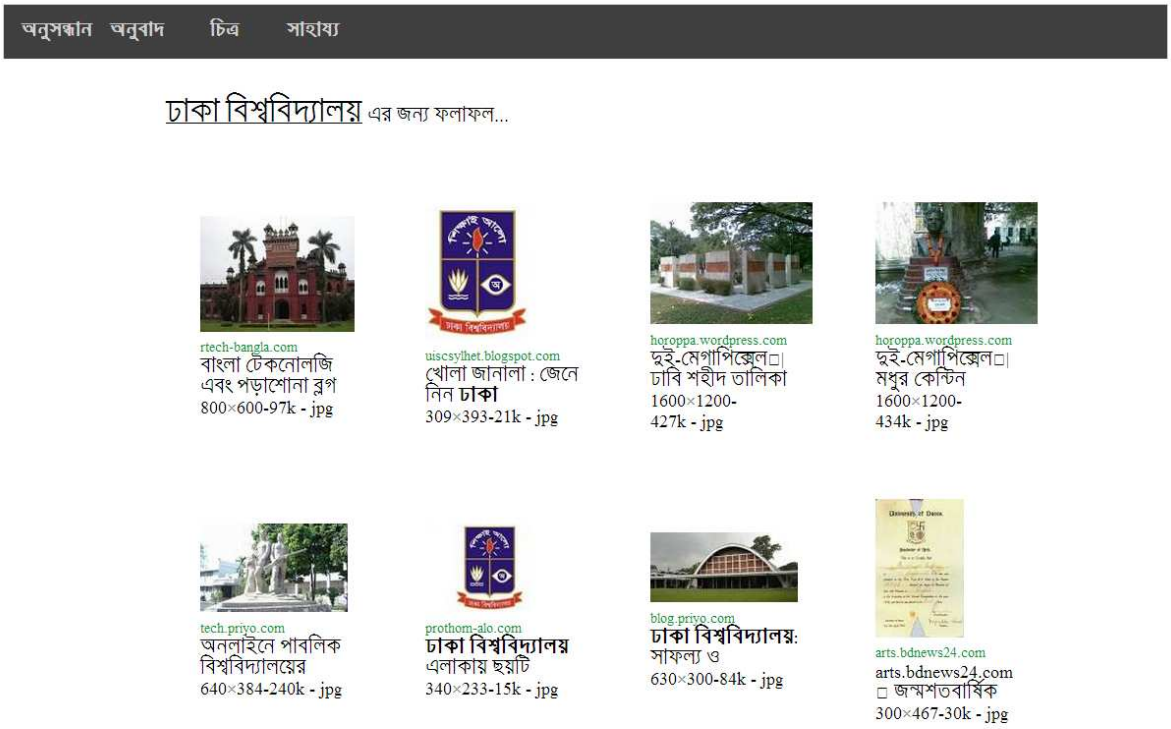
চাকরি > চাকরি ঢাকা বিভাগ - Buy & Sell easily at Bikroy.com
bikroy.com/bn/jobs-in-dhaka-division
বিভিন্ন পেশা এবং চাকরির সুযোগ **ঢাকা বিভাগ**।বিজ্ঞাপন খুঁজে **...**

ঢাকা বিভাগ | ঢাকা বিভাগ
www.dhakadiv.gov.bd/
প্রথম পাতা. বিভাগ সম্পর্কিত. বিভাগ পরিচিতি. বিভাগের পটভুমি; এক নজরে **ঢাকা বিভাগ**; মানচিত্রে বিভাগ; জেলা ও উপজেলা. ইতিহাস ও ঐতিহ্য. বিভাগের ঐতিহ্য; পুরাকীর্তির সংক্ষিপ্ত বর্ণনা; মুক্তিযুদ্ধ ও মুক্তিযোদ্ধা; প্রখ্যাত ব্যক্তিত্ব; ভাষা ও সংস্কৃতি; খেলাধুলা ও বিনোদন. ভৌগলিক ও অর্থনৈতিক. প্রাকৃতিক সম্পদ; নদ নদী; ব্যবসা বাণিজ্য **...**

FIGURE 4.2: Results for Web Search

After learning about the token-type pair the system enters a new entry of that token-type in the knowledge graph. When next time some other user enters that token as search key, he or she can have suggestion information about it.

4.5 Summary



ঢাকা বিশ্ববিদ্যালয় এর জন্য ফলাফল...

rtech-bangla.com
বাংলা টেকনোলজি এবং পড়াশোনা ব্লগ
800×600-97k - jpg

uiscsylhet.blogspot.com
খোলা জানালা : জেনে নিন **ঢাকা**
309×393-21k - jpg

horoppa.wordpress.com
দুই-মেগাপিক্সেল□| ঢাবি শহীদ তালিকা
1600×1200-427k - jpg

horoppa.wordpress.com
দুই-মেগাপিক্সেল□| মধুর কেন্টিন
1600×1200-434k - jpg

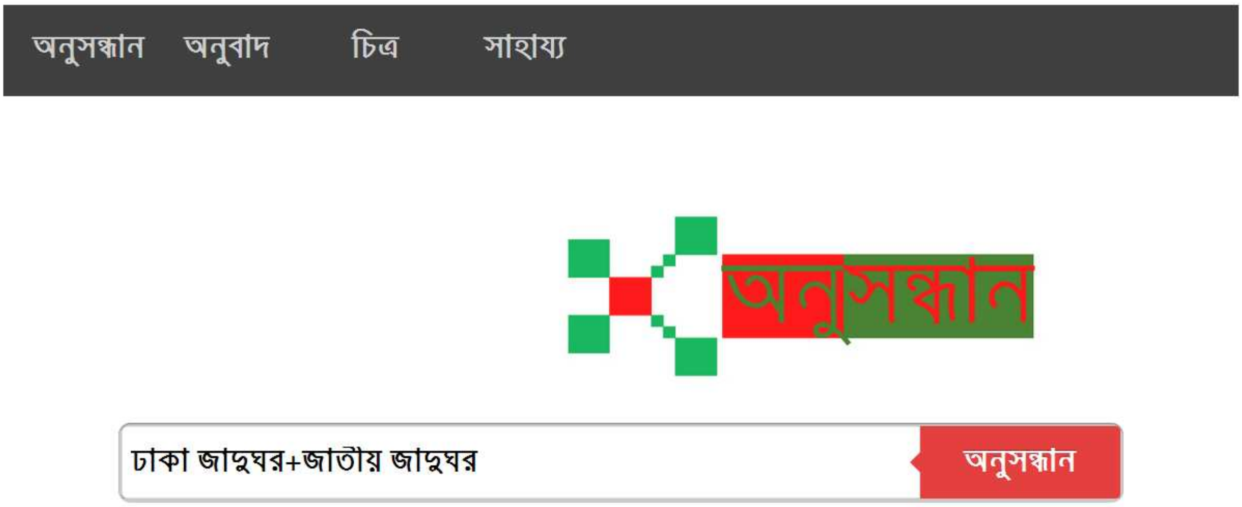
tech.priyo.com
অনলাইনে পাবলিক বিশ্ববিদ্যালয়ের
640×384-240k - jpg

prothom-alo.com
ঢাকা বিশ্ববিদ্যালয় এলাকায় ছয়টি
340×233-15k - jpg

blog.priyo.com
ঢাকা বিশ্ববিদ্যালয়: সাফল্য ও
630×300-84k - jpg

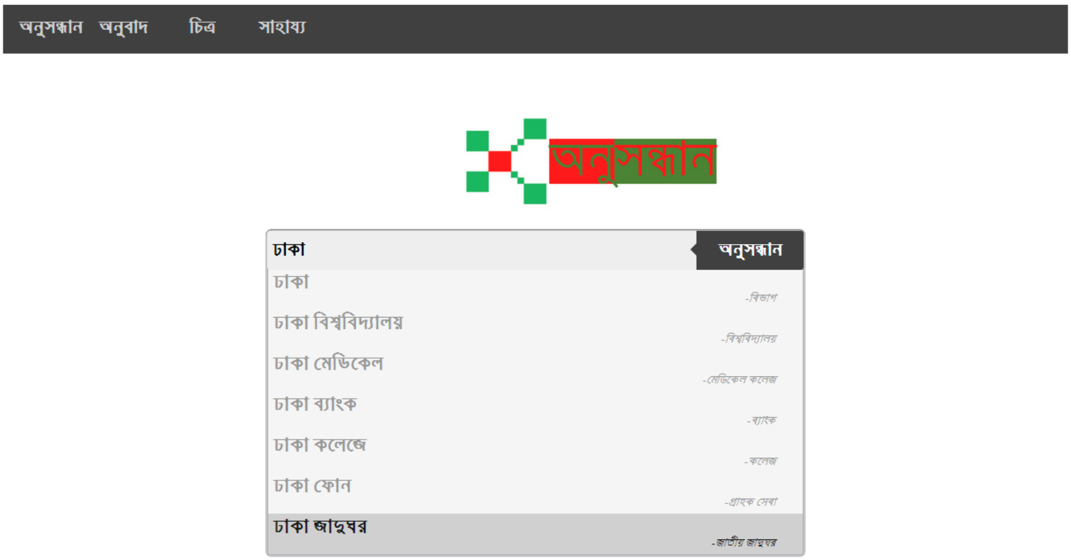
arts.bdnews24.com
arts.bdnews24.com □ জন্মশতবার্ষিক
300×467-30k - jpg

FIGURE 4.3: Results for Image Search



ঢাকা জাদুঘর+জাতীয় জাদুঘর

Figure 4.4: Learning from User



ঢাকা

ঢাকা -বিভাগ

ঢাকা বিশ্ববিদ্যালয় -বিশ্ববিদ্যালয়

ঢাকা মেডিকেল -মেডিকেল কলেজ

ঢাকা ব্যাংক -ব্যাংক

ঢাকা কলেজে -কলেজ

ঢাকা ফোন -গ্রাহক সেবা

ঢাকা জাদুঘর -জাতীয় জাদুঘর

FIGURE 4.5: Learnt Knowledge

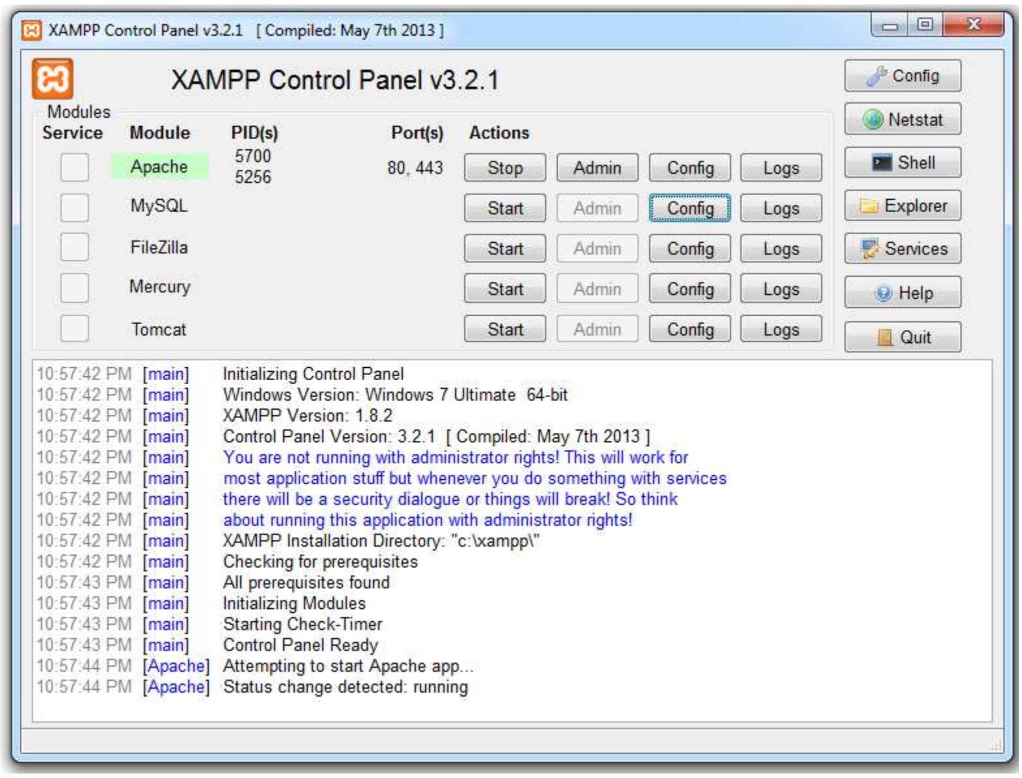
Chapter 5

Design Tools

5.1 Xamp Server

XAMPP is a free and open source web server package. It has an all-in-one installer that installs the entire package in a single directory. This enables users to use them without further modification. The XAMPP server mainly consists of the Apache HTTP Server, MySQL database, and interpreters for scripts written in the PHP and Perl programming languages. It provides an interface called PhpMyAdmin that can be used to create and maintain databases graphically. The services provided by the server are

- Apache
- MySQL
- FileZilla
- Mercury
- Tomcat



```
10:57:42 PM  [main]    Initializing Control Panel
10:57:42 PM  [main]    Windows Version: Windows 7 Ultimate  64-bit
10:57:42 PM  [main]    XAMPP Version: 1.8.2
10:57:42 PM  [main]    Control Panel Version: 3.2.1  [ Compiled: May 7th 2013 ]
10:57:42 PM  [main]    You are not running with administrator rights! This will work for
10:57:42 PM  [main]    most application stuff but whenever you do something with services
10:57:42 PM  [main]    there will be a security dialogue or things will break! So think
10:57:42 PM  [main]    about running this application with administrator rights!
10:57:42 PM  [main]    XAMPP Installation Directory: "c:\xampp\"
10:57:42 PM  [main]    Checking for prerequisites
10:57:43 PM  [main]    All prerequisites found
10:57:43 PM  [main]    Initializing Modules
10:57:43 PM  [main]    Starting Check-Timer
10:57:43 PM  [main]    Control Panel Ready
10:57:44 PM  [Apache]  Attempting to start Apache app...
10:57:44 PM  [Apache]  Status change detected: running
```

Figure 5.1: XAMPP Server

5.2 NetBeans IDE

NetBeans is an integrated development environment (IDE) for developing with languages like Java, PHP, C/C++, and HTML5. Maintaining projects in NetBeans is very simple and efficient. It provides many build-in language supports. Also it can be used as testing and debugging tool.

5.3 PHP

PHP is a server-side scripting language. It was mainly designed for web development but it can also be used as general-purpose programming language. Web pages are generated dynamically via interpreting PHP code with the web server. PHP commands can be embedded directly into an HTML source document rather than calling an external file to process data. Runtime compilation of scripts by PHP engine enhances execution speed. PHP has a lot in common with most of the higher level language when it comes to the flow control statements and loop statements. It also supports most of the primitive data types.

5.4 Ajax

Ajax stands for Asynchronous JavaScript and XML. It is a group of interrelated web development techniques used on the client-side to create asynchronous web applications. With Ajax, web applications can send data to, and retrieve data from, a server asynchronously without interfering with the display and behavior of the existing page. It is generally used when fast communication between the server and the client is needed without interrupting the view on the server site.

Chapter 6

Conclusion

6.1 Conclusion

Appendix A

Appendix Title Here

Write your Appendix content here.

Bibliography