

# **Knowledge Classification Agent**

**By**

**Faliq Haffiz Alawi**

*Dissertation submitted in partial fulfillment of  
the requirement for the*  
Bachelor of Technology (Hons)  
(Information System)

**JULY 2005**

**UNIVERSITI TEKNOLOGI PETRONAS  
BANDAR SERI ISKANDAR 31750  
TRONOH, PERAK  
JULY 2005**

# **CERTIFICATION OF APPROVAL**

## **KNOWLEDGE CLASSIFICATION AGENT**

By

Faliq Haffiz Alawi

(3208)

A project dissertation submitted to the

Information System Programme

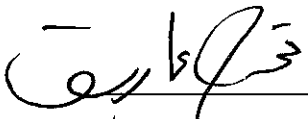
University of Technology PETRONAS

In partial fulfillment of the requirement for the

BACHELOR OF TECHNOLOGY (Hons)

(INFORMATION SYSTEM)

Approved by,



(Pn. Mazeyanti Mohd Arifin)

UNIVERSITI TEKNOLOGI PETRONAS

TRONOH, PERAK

JULY 2005

## **CERTIFICATION OF ORIGINALITY**

This is to certify that I am responsible for the work submitted in the project, that the originality work is my own except as specified in the references and acknowledgment and that the original work contain here have not been undertaken or done by unspecified sources or person.



(Faliq Haffiz Alawi)

## **ABSTRACT**

This paper is about the knowledge classification agent. This system is a knowledge management based that integrated the way of searching information via internet using a search engine. The availability of a search engine especially in Malaysia is not as efficient as it can be. Using a normal search engine, the search result is too general and users usually get results which are not always suit their request. This system will use an agent that work by grouping the result into related groups or categories. Before producing the final results to the user, this system will categorize the result into common type of keyword, say, user searching a keyword 'beetles' this agent will group it into groups such beetles for cars, beetles for insects and beetles for music band. The target users of the system are people that use the internet as an information resource such as academicians and researchers. Useful information that had been classified needed by users in order to choose the best information is identify as the main problems that trigger the project. The objectives of the project are to develop classified results into related groups for users so that users can find the requested information efficiently as well as save users time. Author had planned to use a spiral model as a methodology. Author believes that by using this system, the problems that stated above can be solved.

## ACKNOWLEDGEMENT

Alhamdulillah, praised to be to Allah SWT, God of the universe, and peace be upon Prophet Muhammad SAW the Messenger of Allah. With the permission of Allah and His Bless I may complete my Final Year Project (FYP) almost successfully. Hereby I would like to take this opportunity to convey my deepest appreciation to all people that involve throughout the project development whether it is directly or indirectly. Their contributions to my project are really meaningful.

First and for most, I would like to thank my supervisor Pn.Mazeyanti Mohd Arifin at the department of Information System, University of Technology PETRONAS for her guidance and patient throughout the development of this project. Without her help I may not be able to complete all my task even though I could not completed my project as it suppose.

Thanks to Ms Nazleeni, Pn Syarifah Bahiyah and Ms Noreen Izza at the Department of Information System for their technical advice and guidance. Not forgotten, Mohd Yaakup (Dun), ZD, Anuar Fariz Jameli and Bob Lavau for their support and help in the project development. This acknowledgement also dedicated to all my housemate Mohd Amri Azmi, Mohd Hafriz Roslee, Mohd Shahrul Anuar Ishak, Mohd Khairuddin Adnan, Muhd Faisel Amin, Khairul Anuar Abu Bakar, Che Wan Kamril Nisyam Che Wan Awang, Mohd Syfizan Dato' Shahir and Mohd Iskandar Yaakob for their help in doing my test on the dummy system and as well as for their support and affection.

Greatest appreciation to my family patiently waits and prays for my success. Special thanks to Khairul Alia Khairuddin even when she is not part of this project but she will always be a part of me.

Finally, thanks to all my friends and anybody who I may not mention their name. Their help and support may no be forgotten. For now, Wallahua'lam. Thank you very much and may Allah Bless all of you.

## TABLE OF CONTENT

CERTIFICATION OF APPROVAL	. . . . .	ii
CERTIFICATION OF ORIGINALITY	. . . . .	iii
ABSTRACT	. . . . .	iv
ACKNOWLEDGEMENT	. . . . .	v
TABLE OF CONTENT	. . . . .	vi
 <b>CHAPTER 1: INTRODUCTION</b>	 . . . . .	 <b>1</b>
1.1 Background of study	. . . . .	2
1.2 Problem Statements	. . . . .	2
1.3 Problem Identification	. . . . .	2
1.4 Significant of the Project	. . . . .	2
1.5 Objectives	. . . . .	2
1.6 Scope of Studies	. . . . .	3
 <b>CHAPTER 2: LITERATURE REVIEW AND THEORY</b>	 . . . . .	 <b>4</b>
2.1 Researched on googlebot	. . . . .	4
2.2 Googlebot Results	. . . . .	10
2.3 Classifier	. . . . .	12
2.4 Web Crawler Works	. . . . .	15
 <b>CHAPTER 3: METHODOLOGY / PROJECT WORK</b>	 . . . . .	 <b>16</b>
3.1 Procedure Identification	. . . . .	17
3.2 User's Communication	. . . . .	18
3.3 Risk Analysis	. . . . .	18
3.4 Engineering	. . . . .	18
3.5 Construction	. . . . .	19
3.6 User's Evaluation	. . . . .	19

**CHAPTER 4: RESULT AND FINDINGS . . . 21**

4.1 Main Concept . . . 22

4.2 Hardware Requirement . . . 23

**CHAPTER 5: CONCLUSION AND RECOMMENDATION 24**

**REFERENCES . . . 25**

# **CHAPTER 1**

## **INTRODUCTION**

### **1.1 Background of study**

The increasing numbers of websites resulted from the search engine say, google.com or yahoo.com made it inefficient for users wants as resulted from the survey that had been conducted by the author (see appendices). All the results were listed according to keyword that users' entered for searching the information needed. The results displayed were unclassified and users commonly have to search it one by one to find the information that they need. Ergo, users find it is inefficient to search the information using classic internet search engine. Realizing this problem, author had proposed a system called Knowledge Classification Agent.



## **1.2 Problem Statement**

### **1.2.1 Problem Identification**

Classifications were needed in the result of a search engine because:

- Less efficiency of one search engine
  - The results were too general in which it is not always what the user's requested for.
  - The results performed by a search engine were normally unclassified.

### **1.2.2 Significant of the Project**

The benefits of this project are:

- As an easy way for users to find requested information resulted from classified results from a search engine.
- Encourage offspring's and so the millennium society to use the internet as information resources as using author's project will make the search job more efficient.

### **1.3 Objectives**

The objectives of the project are as follows:

- To classify all the results from the search engine to be precisely fits the users requested information.
- To reduce time.

### **1.4 Scope of studies**

The scope of studies throughout the project includes:

- Classifying results taken from a search engine. Involving studies on knowledge management (KM) where the system can produce a better way on searching information.
- Develop a system that involves classifying using intelligent agent.

## **CHAPTER 2**

### **LITERATURE REVIEW**

Knowledge management (KM) is a methodology in which it is new here in Malaysia. It is a management of knowledge between organizations. It helps a person to make the best decision. Before since, decisions are made by paperwork researched that cause a person time constraint and headache. According to A. Tiwana, the Knowledge Management Toolkit: Orchestrating IT, Strategy, and Knowledge Platforms (2nd Edition), Upper Saddle River, NJ: Prentice Hall, 2002.

Knowledge management is a widely accepted 'working definition' of knowledge management applied in worldwide organizations is available from the:

"Knowledge Management caters to the critical issues of organizational adoption, survival and competence in face of increasingly discontinuous environmental change. Essentially, it embodies organizational processes that seek synergistic combination of data and information processing capacity of information technologies, and the creative and innovative capacity of human beings." [6]

Thus, this supported information gave the author a bit of information on why KM is needed.

## **2.1 Research on googlebot built by google.com**

Frequently asked question according to surveys. (Surveys was conducted by google.com)

### **1. How often will Googlebot access my web pages?**

- For most sites, Googlebot shouldn't access your site more than once every few seconds on average. However, due to network delays, it's possible that the rate will appear to be slightly higher over short periods.

### **2. How do I request that Google not crawl parts or my entire site?**

- Robots.txt is a standard document that can tell Googlebot not to download some or all information from your web server. The format of the robots.txt file is specified in the Robot Exclusion Standard. For detailed instructions about how to prevent Googlebot from crawling all or part of your site, please refer to our Removals page. Remember, changes to your server's robots.txt file won't be immediately reflected in Google; they'll be discovered and propagate when Googlebot next crawls your site.

### **3. Googlebot is crawling my site too fast. What can I do?**

- Please contact us with the URL of your site and a detailed description of the problem. Please also include a portion of the weblog that shows Google accesses so we can track down the problem quickly.

### **4. Why is Googlebot asking for a file called robots.txt that isn't on my server?**

- Robots.txt is a standard document that can tell Googlebot not to download some or all information from your web server. For information on how to create a robots.txt file, see The Robot Exclusion Standard. If you just want to prevent the "file not found" error messages in your web server log, you can create an empty file named robots.txt.

5. Why is Googlebot trying to download incorrect links from my server? Or from a server that doesn't exist?

- It's a given that many links on the web will be broken or outdated at any particular time. Whenever someone publishes an incorrect link to your site (perhaps due to a typo or spelling error) or fails to update links to reflect changes in your server, Googlebot will try to download an incorrect link from your site. This also explains why you may get hits on a machine that's not even a web server.

6. Why is Googlebot downloading information from our "secret" web server?

- It's almost impossible to keep a web server secret by not publishing any links to it. As soon as someone follows a link from your "secret" server to another web server, your "secret" URL may appear in the referrer tag and can be stored and published by the other web server in its referrer log. So, if there's a link to your "secret" web server or page on the web anywhere, it's likely that Googlebot and other web crawlers will find it.

7. Why isn't Googlebot obeying my robots.txt file?

- To save bandwidth, Googlebot only downloads the robots.txt file once a day or whenever we've fetched many pages from the server. So, it may take a while for Googlebot to learn of changes to your robots.txt file. Also, Googlebot is distributed on several machines. Each of these keeps its own record of your robots.txt file.
- We always suggest verifying that your syntax is correct against the standard at <http://www.robotstxt.org/wc/exclusion.html#robotstxt>. A common source of problems is that the robots.txt file isn't placed in the top directory of the server.
- Also, there's a small difference between the way Googlebot handles the robots.txt file and the way the robots.txt standard says we should (keeping in mind the distinction between "should" and "must"). The standard says we should obey the first applicable rule, whereas Googlebot obeys the longest (that is, the most specific) applicable rule. This more intuitive practice matches what people actually do, and what they expect us to do. For example, consider the following robots.txt file:

User-Agent: \*

Allow: /

Disallow: /cgi-bin

- It's obvious that the webmaster's intent here is to allow robots to crawl everything except the /cgi-bin directory. Consequently, that's what we do.

8. Why are there hits from multiple machines at Google.com, all with user-agent Googlebot?

- Googlebot was designed to be distributed on several machines to improve performance and scale as the web grows. Also, to cut down on bandwidth usage, we run many crawlers on machines located near the sites they're indexing in the network.

9. Can you tell me the IP addresses from which Googlebot crawls so that I can filter my logs?

- The IP addresses used by Googlebot change from time to time. The best way to identify accesses by Googlebot is to use the user-agent (Googlebot).

10. Why is Googlebot downloading the same page on my site multiple times?

- In general, Googlebot should only download one copy of each file from your site during a given crawl. Very occasionally the crawler is stopped and restarted, which may cause it to recrawl pages that it's recently retrieved.

11. Why don't the pages of my site that Googlebot crawled show up in your index?

- Don't be alarmed if you can't immediately find documents that Googlebot has crawled in the Google search engine. Documents are entered into our index soon after being crawled. Occasionally, documents fetched by Googlebot won't be included for various reasons (e.g. they appear to be duplicates of other pages on the web).

12. What kinds of links does Googlebot follow?

- Googlebot follows HREF links and SRC links.

13. How do I prevent Googlebot from following links on my pages?

- To keep Googlebot from following links on your pages to other pages or documents, you'd place the following meta tag in the head of your HTML document:

```
<META NAME="Googlebot" CONTENT="nofollow">
```

- To learn more about meta tags, please refer to <http://www.robotstxt.org/wc/exclusion.html#meta>; you can also read what the HTML standard has to say about these tags. Remember, changes to your site won't be immediately reflected in Google; they'll be discovered and propagate when Googlebot next crawls your site.

14. How do I tell Googlebot not to crawl a single outgoing link on a page?

- Meta tags can exclude all outgoing links on a page, but you can also instruct Googlebot not to crawl individual links by adding `rel="nofollow"` to a hyperlink. When Google sees the attribute `rel="nofollow"` on hyperlinks, those links won't get any credit when we rank websites in our search results. For example a link,

```
<a href=http://www.example.com/>This is a great link!</a>
```

could be replaced with

```
<a href=http://www.example.com/ rel="nofollow">I can't vouch for this  
link</a>.
```

15. What is Feedfetcher, and why is it ignoring my robots.txt file?

- Feedfetcher requests come from explicit action by human users. When users add your feed to their Google homepage or to Google Reader, Google's Feedfetcher attempts to obtain the content of the feed in order to display it. Since all requests come from humans, Feedfetcher has been designed to ignore robots.txt. Learn more.

16. How do I add my feed to the search results for Google's personalized homepage and Google Reader?

- The feeds that Googlebot crawls appear in the search results for Google's personalized homepage and Google Reader. To ensure that your feed is part of this index, add a <link> tag to the header of your webpage to enable feed autodiscovery. There are a lot of variations on <link> tags for this purpose, but below are a couple simple examples.

For an Atom feed:

```
<link rel="alternate" type="application/atom+xml" title="Your Feed Title"
href="http://www.example.com/atom.xml" />
```

- For an RSS feed:

```
<link rel="alternate" type="application/rss+xml" title="Your Feed Title"
href="http://www.example.com/rss.xml" />
```

//taken from [www.google.com/googlebot](http://www.google.com/googlebot) [8]

According to all these questions and answers, author had found out how the crawler worked. What the crawler does and how to prevent crawler such 'googlebot' to crawl one website. In all of this question and answers, this related information is to be use as author project is dealing with finding the result and then group them into groups.



## 2.2 These are the details that googlebot provide after crawling one site

### The Web Robots Pages

#### Googlebot

Name	: Googlebot
Cover Page	: <a href="http://www.googlebot.com/">http://www.googlebot.com/</a>
Details Page	: <a href="http://www.googlebot.com/bot.html">http://www.googlebot.com/bot.html</a>
Operational Status	: active
Description	: Google's crawler
Robot Purpose	: indexing
Software Type	: standalone
Software Platform	: Linux
Software Language	: c++
Availability	: none
Owner's Name	: Google Inc.
Owner's Home Page	: <a href="http://www.google.com/">http://www.google.com/</a>
Owner's Email Address:	googlebot@google.com
Exclusion Protocol	: yes
Exclusion Tag	: googlebot
Supports NOINDEX	: yes
Robot Host	: googlebot.com
HTTP From	: yes
HTTP User-Agent	: Googlebot/2.X (+ <a href="http://www.googlebot.com/bot.htm">http://www.googlebot.com/bot.htm</a> )
History	: Developed by Google Inc
Environment	: commercial
Identifier	: googlebot
Updated	: Thu Mar 29 21:00:07 PST 2001
Update By	: googlebot@google.com

\*taken from googlebot. [7]

## 2.3 Classifier

The main thing to be developed is the classifier, a classifier that can classify the results from one search engine. As of that, author had also done some research on how to classify, or group the result.

Beginning with Thunderbird version 0.9, the Message Grouping (Group by Sort) feature lets you organize messages in the message list into self-contained groups according to attributes such as date, sender or priority. For instance, if you set your Inbox to group messages by date, the messages will be organized into folder-like groups labeled with today, yesterday, last week, and and so on. You can expand or collapse each group by clicking on the small +/- mark next to the group label. Messages can be grouped according to date, priority, sender, recipient, status, subject, or label.

To group messages: first click on the folder that you want to use (such as your Inbox) and sort the messages in your preferred way: either click on a column heading such as "Date" in the message-list pane, or make a selection from "View -> Sort by". Then select "View -> Sort by -> Grouped by Sort" or simply press the letter "G" on your keyboard.

To turn off message grouping: click on any column heading in the message-list pane or choose a different sort order from "View -> Sort by". Other information Message grouping works on a per-folder basis. For example, message grouping that you apply to the Inbox folder will not be applied to the Sent folder or any other folders. Message grouping does not currently work with Saved Search folders.

According to the above explanation, author had found that classifying can be made in many ways. The classified results can be grouped according to categories such cars, insect, and many other categories. It is possible to classify everything into all categories and it needed some time to make it happen.

## **2.4 General tips on building a classifier**

Search engines do not index all the documents available on the Web. For example, most engines cannot index files to password-protected sites. Good examples of this are the research databases and e-journals licensed for use by libraries and made available to affiliated users only. Documents behind a firewall will not be accessible to a search engine spider. Other files can be excluded from search engines by Web server software at the host site, or by a command within the file itself. Still other Web pages may not be picked up if they are not linked to other pages, and are therefore missed by a search engine spider as it crawls from one page to the next. Search engines rarely contain the most recent documents posted to the Internet; do not look for yesterday's news on a search engine (unless, of course, it offers a separate news search).

The content of databases generally will not show up in a search engine result. This is because search engine spiders cannot or will not get inside database tables and extract the data. The phenomenon is sometimes referred to as the deep Web. Later on in this tutorial, we will examine the nature of the deep Web.

Search engine features are proliferating and are in a constant state of flux. Don't try to keep track of everything! For example, several search services offer searches on various fields, programming languages, domain locations, dates, and so on. As search engines develop and the competition among them intensifies, more features are available to users in all sorts of combinations. For a review of some of these features, and the engines that support them, see [How to Choose a Search Engine or Research Database](#).

Most major search engine indexes consist of the full text of source files. When you search a full text index, you will retrieve a file even if your search terms appear only once in the text and do not represent the primary topic of the document. (Of course,

a good engine will place this type of file low in the list of results based on its relevancy ranking scheme.) Limiting your search to fields or using proximity operators (explained below) can be a useful way to boost the relevancy of your results.

Many search engines have an interface for basic searches as well as a separate interface for advanced or more full-featured queries. Be sure to explore both interfaces and to use the one that is appropriate for your query. Keep in mind that some advanced search interfaces may actually be easier to use than the interface on the main screen. Visit AltaVista for an example of this.

Because of the potentially large number of pages that can be retrieved by a search, good relevancy ranking is important. Most search engines use various criteria to construct a relevancy rating of each hit and will present your search results in this order. First generation search engines primarily use term relevancy ranking. This type of ranking judges relevancy based on the presence of your search terms in Web documents. For example, ranking will be based on: the presence of search terms in the title, URL, first heading; the number of times search terms appear in the document; search terms appearing early in the document; search terms appearing close together; etc.

This is known as "on the page" ranking, since the engine looks at content on the page to determine its relevancy. The use of "on the page" ranking as the sole ranking scheme has been fading from the search engine scene because it has proven to be too simplistic for the Web environment.

One of the most interesting developments in search engine technology is the organization of search results by peer ranking and the bundling of results into component concepts, domains and sites in addition to term relevancy. This type of ranking looks at "off the page" information to determine the order of your search results. Search engines that employ this alternative may be thought of as second generation search services. For example:

- Google ranks by the number of links from pages ranked high by the service

- Teoma ranks according to the number of links from topically relevant pages
- Vivisimo sorts results into categories representing concepts derived from your search

A more detailed look at second generation search services may be found in the tutorial *Second Generation Searching on the Web*.

Search tools generally present results in one of two ways:

- Vertical layout: Your results are presented in one long list. This is by far the most common method of presentation. In these cases, you need to examine each source to determine if it addresses aspects of your topic that interest you.
- Horizontal layout: Certain concept grouping engines offer results in a horizontal layout. With this feature, you can first review concept categories retrieved by your search before examining the results within particular categories. This type of organization can make it easier to determine if your results relate to the aspect of the topic that interests you. Examples of these tools are Query Server and Vivisimo

Don't be impressed by a large number of hits in response to a search. Often multiple pages are returned from a single site because they all contain your search terms. AltaVista is one search engine that avoids this with a technique called results grouping, whereby all the results from one site are clustered together into one result. You are then given the opportunity to view all the retrieved pages from that site if you choose. With these engines, you may get a smaller number of results from a search, but each result is coming from a different site.

Offered features do not always work perfectly. Don't look for perfection. Just relax and get what you can out of the search.

It is helpful to understand that not all aspects of search engine technology are revealed to the public. In the world of commercial search engines, trade secrets abound. Help files tend to be general in nature when explaining how the technology

works. This writer has queried services via e-mail for more details, only to get back slightly more substantive information.

Watch for converging content. Many well-known sites now contain information from an array of sources. Some have applied the term "portal" to describe this phenomenon. Offerings on a portal can increase the usefulness of search sites, but also can create confusion in terms of the information source. For example, consider what you may find on a commercial search engine service:

- Spider gathered index: The mechanism for searching a spider-gathered index is the feature people usually associate with a search engine.
- Results from other search services: It is increasingly common for a search engine to return results from other services with which it has partnered. Each partner service offers an enhancement over search results that are derived from the Web. This represents an interesting combination of first and second generation search technologies appearing on the same site.
- Directory: Many search services offer a directory on their sites. This directory may be a name brand such as LookSmart or the Open Directory Project, or a directory compiled by a site's own editors. Results from the directory may appear automatically with results from the spider-crawled Web, or the directory may be searched or browsed separately.
- Deep Web: Many search services offer the option to search databases offering specific content. Included may be news, business, shopping, multimedia files, and so on. These databases constitute a small subset of the deep Web.

All of this points to a blurring of the distinctions among sites that provide directory content, those that offer results from the spider-crawled Web, and those that provide access to content on the deep Web.

After researched through all the words above, author had found out that a good classifier needs an intelligent agent to classify. This can made the classifier to give more useful results in information searching. With a good search using the crawler or spider such as googlebot can made the results more efficient and only the needed information becomes useful.

//taken from <http://library.albany.edu/internet/eng.html> [1]

## 2.5 Learning How to Classify

**Information architecture: learning how to classify** -September 02, 2004

“If you are a knowledge worker, a key skill you require is how to classify content. Classification skills are needed in order to better organize content on your computer, for your emails, and for how you compose documents. If you have responsibility for a website, classification is an essential skill.”

-By Gerry McGovern

Classification (taxonomy) is a type of metadata. The purpose of metadata is to provide essential information about a document. Metadata and classification are part of the discipline of information architecture, whose focus is to organize and layout content.

Classification is not simple. Classifying 20 documents isn't difficult. That's because no matter how you classify them, it will be relatively easy to find what you want. Classifying 2,000 documents is a very difficult task.

Classification is not something that you can master in a weekend. It will take you years to become expert at classification. However, if you want to master content you must master classification.

Classification is an inherent part of creating a document. Every time you write a heading, you are in fact creating a classification. If the document is long (more than

600 words) you should have sub-headings. These are sub-classifications underneath the heading classification.

Good internal classification has three key objectives:

- To organize the document in such a way that maximizes its ability to communicate knowledge.
- To allow the reader to quickly find specific parts of the document.
- To allow the reader to extract specific parts of various documents, and in so doing create a new document. For example, the reader might compile the summaries of ten documents dealing with the European car industry. (Extensible Mark-up Language (XML) is useful for this sort of task.)

Classification experts tend to focus on organizing complete documents, books, music and other content. They classify for two reasons:

- To organize the content so that it can be found quickly.
- To place the content in context so that it becomes part of a cohesive body of knowledge.



## **CHAPTER 3**

### **METHODOLOGY**

#### **3.1 Procedure Identification**

In developing this project, author planned to use the spiral model. This evolutionary system model that couple the iterative nature of prototyping with the controlled and systematic aspects of the linear sequential model.

The spiral model also provides the potential for rapid development of incremental versions of the system. Using this methodology, the author's project can be develop in a series if incremental releases.

A spiral model is divided into numbers of framework activities, also called task regain. Typically there are between 3 to 6 task regions. The methodology of the project is shown in figure 1

**Project Methodology**

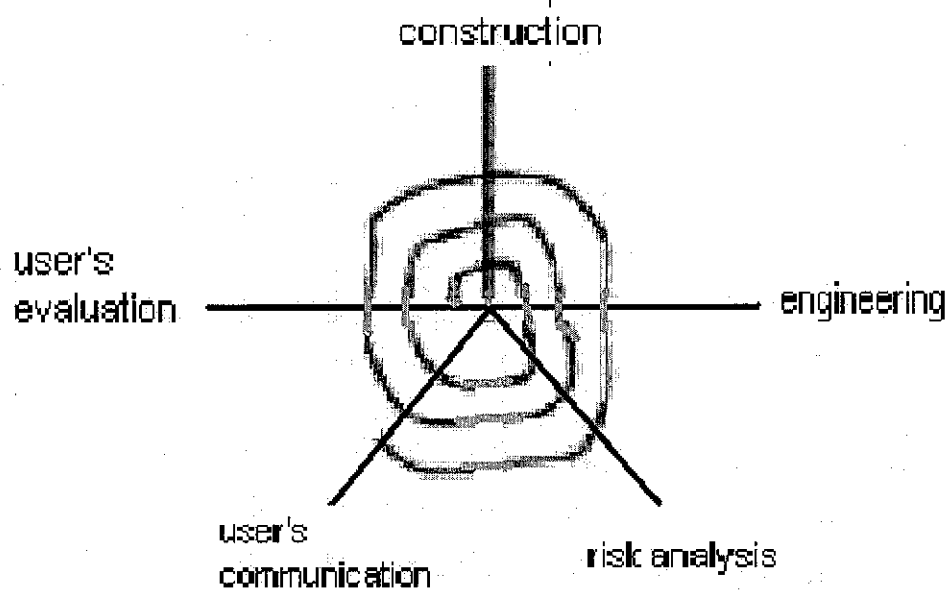


Figure 1: Spiral model

### **3.1.1 Users communication and planning**

The purpose of this phase is to establish effective communication between author and users. This task will be done by interviewing author's colleges. Several fellow friends come to help with joining the interview and answered the questionnaire conducted by the author.

In this phase it's include determine the project background, the problem that trigger the project, the benefit of the project and scope of the project. Some research was done to get more idea and clearer picture on the project. The research focused on journals and articles that related to this project. Other references such as book and websites were also be used as references sources.

### **3.1.2 Risks analysis**

During this phase, author will consider all the problems that might occur during project development. The risk that might occur such accuracy of the author to make the project to become as close as according to the objectives, the ability of the author to conclude and finish the project development within the time limit and if the project is possible to build or not.

### **3.1.3 Engineering**

Representations of the system were developed. The first representation was a meta tag crawler. It is closely like a dummy system but it is a system that represents how results being grouped. This dummy system has a database called 'si agent' whereby 'Si agent' stored data that related to one site. This is where author had presented the system as version 1 or volume 1. Another version of crawler is yet to come but under some circumstances author faces some problem just like what author stated in the risk analysis phase.

### **3.1.4 Construction**

Testing was done during this phase. Testing was conducted throughout preEDX which the result was not as author hoped. As this is a spiral methodology, author believes that after completed the sixth region, in which to start a new round to complete a spiral diagram, this project development can reach its goal.

### **3.1.5 Users' evaluation (Feedback)**

Feedback by the users based on evaluation of the system. Users (author's friends) had been given the opportunity to use the system and with respects gave their opinion about this project. Their feedback was used to find faulty of author's project.

### **3.1.6 Users' communication**

Start the new loop of region. Whereby all the feedbacks will be used as the new requirements to develop enhance agent.

## 3.2 Tools

To develop “knowledge classification agent” the main tools are:

- Dreamweaver
- Java jdk 1.5
- NetServer.

### 3.2.1 The platform and Display project purpose

The tool that had been used to create the platform is Macromedia Dreamweaver mx 4.0. was used because:

- It allows the internet application to be built in it.
- At the same time it can be connected with other tools to be used such Java jdk 1.5 and NetServer.

### 3.2.2 Web Crawler and grouping result

The main tool or software that been used for this purpose is Java jdk 1.5

Java jdk 1.5 software provides functions and tools that needed in order to build the crawler and group the crawled result. This software was developed by Java.

### 3.2.3 Storage and display purpose

Using NetServer Manager 0.1Beta3-Win NT/2000/XP, it provides function to store data directly and perform display over author’s computer. [4]

## **CHAPTER 4**

### **RESULT AND FINDING**

In this chapter, the author had written the result and discussion for the analysis. This chapter focuses on the concept and system architecture, process taken to develop the project.

#### **4.1 Main concept**

As described in the previous chapter, the architecture of knowledge classification agent consists of grouping the search result into possible categories.

Firstly, the agent will take out all the results from the search engine resulted from what user required. For instant, user request to find information about beetle, the search results will go through the agent to be classified.

Secondly, the agent classifier will then classifies the results according to the possible categories such beetle for insect, beetle for car, beetle for music band and any category that the keyword beetle will fall to.

Lastly, all the classified results will be display for the user to be chosen which category the user requested for. The user has to click on the link available to get the information. The link is the URL to the site that has the information the user's wants.

## 4.2 Build a Crawler

To build a crawler that can crawl the internet, here's the pseudocode of the algorithm:  
Get the user's input: the starting URL and the desired file type. Add the URL to the currently empty list of URLs to search. While the list of URLs to search is not empty,

```
{  
    Get the first URL in the list.  
    Move the URL to the list of URLs already searched.  
    Check the URL to make sure its protocol is HTTP  
    (if not, break out of the loop, back to "While").  
    See whether there's a robots.txt file at this site  
    that includes a "Disallow" statement.  
    (If so, break out of the loop, back to "While".)  
    Try to "open" the URL (that is, retrieve that document From the Web).  
    If it's not an HTML file, break out of the loop,  
    back to "While."  
    Step through the HTML file. While the HTML text  
    contains another link,  
    {  
        Validate the link's URL and make sure robots are  
        allowed (just as in the outer loop).  
        If it's an HTML file,  
        If the URL isn't present in either the to-search  
        list or the already-searched list, add it to  
        the to-search list.  
        Else if it's the type of the file the user  
        requested,  
        Add it to the list of files found.  
    }  
}
```

//revision from java.sun [3]

## 4.3 Hardware Requirement

The minimum configuration and requirement that required for supporting application and tools are:

**Basic or Minimum Requirement:**

- CPU : 1.5 GHz Pentium IV
- Memory : 256 MB RAM (recommended)
- Monitor : 14" monitor with 32MB VRAM
- Hard Drives : Hard Drive 10GB
- Media Drives : 40x IDE CD-Rom, 1.44 MB Floppy
- Operating System : Windows 95, 98, ME, NT 4.0, 2000, XP
- Network Card : 10mbps – 100mbps



## **CHAPTER 5**

### **CONCLUSION AND RECOMMENDATION**

Hopefully with the new idea on developing this system will help the users to find that classified information is more useful as well as reduce time for searching information. This new solution can also utilized the usage of search engine by providing grouped results to the user.

By integrating this available solution, this system will go as further to the goal as there will be some one wants to continue this project development. With just a slight bit of researching, the development will go according to the project flow and lastly reached its goal, which is to continue on creating a “Knowledge Classification Agent”.

Hopefully, this research can be continued for the future as this is the agent that can categorize the search engine results into possible categories.

## REFERENCES

- [1] Classifier, [http://kb.mozillazine.org/Message\\_Grouping](http://kb.mozillazine.org/Message_Grouping)
- [2] Crawler, <http://library.albany.edu/internet/eng.html>
- [3] Java Land, <http://java.sun.com/developer/technicalArticles>
- [4] Sun Microsystems Inc, [http://Java.Sun/third part](http://Java.Sun/third%20part)
- [5] NetServer, Sourceforge.net
- [6] According to A. Tiwana, the Knowledge Management Toolkit: Orchestrating IT, Strategy, and Knowledge Platforms (2nd Edition), Upper Saddle River, NJ: Prentice Hall, 2002.
- [7] Googlebot, google.com
- [8] Research on googlebot built on: [www.google.com/googlebot](http://www.google.com/googlebot)
- [9] Gerry McGovern journals Sept 02, 2004.