G.A.S.E.T "TM"

G.A.S.E.T ™ (GETCA Advanced Search Enhancement Technology) ™

An Interactive Search Engine … With Intelligence.

G.A.S.E.T – G

(GETCA Advanced Search Enhancement Technology for Google™) *

* Temporary name : we are using Google™ search engine services for demonstrative and comparison purposes only . All of the Google ™ modified diagrams shown in this document, will be used only as a G.A.S.E.T- G ™ features demonstration tool, bending final approval by the Google ™ Inc legal department. GETCA Inc to the best of its knowledge is following carefully all laws and regulations stated by Google ™ Inc in regard the use of its searching services, and logos.

We will use the name G.A.S.E.T - G when we explain tech Issues, web services, functions and features of G.A.S.E.T project which Is temporarily using Google™ services as a web pages repository.

Contact Us !

Project Dedicated Website : www.gaset-gbset.com

Canada office:

Mr. Gharbiyeh Wael **CEO** / **wael@gaset-gbset.com**
Tel: 647 262 2893
Address: 160 Cactus Ave. # 26 Toronto, Ontario M2R 2V3 (Temporarily)

Jordan office:

Mr. Gharbeyah Wiam **COO** / **weam@gaset-gbset.com**
Tel : 00962 79 673 5642 / 00962 777 460 782
Address : VELA # 28 suliman toqan st / Amman - Jordan

For **General Info** please contact our PR manager at : **info@gaset-gbset.com**
For **Investment Relations** please contact us at : **invest@gaset-gbset.com**
For **Tech Info** please contact our development department : **tech@gaset-gbset.com**

GETCA Inc Strategic Steering Committee (GSSC):

We are very honored at GETCA Inc to have the following Distinguished AI related **Authorities**, **Supervising** our projects (Next Stage) launching. Their Very Impressive academically oriented background in Intelligent systems related fields such as: Smart Web Agents, the **Semantically** Powered / Enhanced Search Engines, Autonomous (**Web Knowledge Utilizing Methods**), and the Innovative **Web Based Chatbots** - Interactive Platforms will Eventually (Guarantee) our project superiority in the search engines market.

- **Dr. Abu Shawar Bayan**
- **Dr. Aldiab Motasem**
- **Dr. Ghnemat Rawan**
- **Dr. Jaber Tareq**

And the kind support of Professor Dr. Khaled El-Zayyat

For more info please check: http://gaset-gbset.com/EXPERTS.html

Part Two

The Implemented Technology !

Table Of Contents / Pages

Introduction

Technical Preview

Description of Main System Components and Techniques

This is a basic / non technical description of some of the technologies and Techniques we developed, more details will be provided upon request .

Problems in Web Searching

- Excellence search results require explicit web **surfing knowledge**, which might take years to acquire. The web searchers will also need **inclusive** language understanding–employing ability .
- Current standard search engines lack the friendly " **Dialogue Based** " **interactive** capabilities, Which will **intelligently** assist the users reaching more productive results.
- Both the Open Web and the Hidden Web are characterized by problems of information coverage, quality, overload, relevancy, currency and completeness, as well as unsuited user interfaces

The Solution … G.A.S.E.T - Beta Version 3.6

- <> **Interactive:** Dialogue-based web search capability with a personal - informative touch
- <> **Intelligent:** AI Powered with Semantically "multitasked - multifaceted" analysis capabilities
- <> **Inclusive:** Searching (multimedia, blogs, books … etc) web resources, all under one roof
- <> **Innovative:** Efficient "Out of the Box " philosophy without "Reinventing the Wheel"
- <> **Implemented:** Partially functional, yet more advanced than standard search engines

Strategic Goals and Planned Course of Action / A

We are currently in the second stage (out of four stages project) of our ongoing mission to build a comprehensive, AI powered search engine with dialogue-based interactive interface (with human-like personal touch), together with the virtual capacity to be (Customized / Enhanced) in harmony with the web surfer level of online search expertise and knowledge – language proficiency.

Our G.A.S.E.T project will boost the current traditional (1990s) "keywords based" web searching philosophy and methods to a higher levels, this will be achieved by applying our experimental AI powered semantically based technologies such as: NLP/G, QA, … etc, and the dynamic utilization of our autonomous web knowledge enhancing techniques.

Strategic Goals and Planned Course of Action / B

The non official 100 + ranking factors used by main search engines like Google ™ to determine results weight (the most single important part of the search engines mythology) was taken in consideration, it was studied - tested extensively by our project strategic steering committee, and even though some of its main points like page rank, html tags, title ... etc where important, still we think it wasn't ambitious enough to match our goals. We think that it leaned too much toward the commercial aspects (current major factors) than the old dictionary simulated ones (early nineties / pre Google ™ techniques), new methods of web pages ranking should be dynamic, multifaceted and depend on the soul of the query, taking into consideration that concepts are time sensitive - evolving objects and people will look at it differently.

Its not enough to give the web surfers choices which we anticipated as result of inflexible formulas, we should be sensitive to the uniqueness of their web surfing habits and the query terms they use, taking into consideration search engines negative practices like: stop words elimination which lead to destruction of info structure, lack of culture sensitivity ... etc.

We also should have a technology which is capable to interact with the web surfers in a friendly intelligent manner, to give them pinpoint results and save them time, we think that we have the needed base for such technology and we can develop our current tools to put into practice.

Strategic Goals and Planned Course of Action / C

We need to continue with our ongoing strategy to build a system that could dynamically - autonomously extract knowledge from the web resources and utilize such knowledge to find semantic, conceptual and contextual relations, leading our system to " learn " from such potentials, it will need to acquire Data in the form of " multifaceted conceptual chunks " which would be:

1. Verified (by comparing it to other stored confirmed or uncertain concepts strings)
2. Enhanced (by making its semantic – contextual blocks design auto upgradeable)
3. linked (with compatible / diverse forms of relations : coordinate, derivative … etc)
4. Evolve (used as a base for suggested evolving concept, context and knowledge)
5. Customizable (to different types of applications labelling : QA, Power Search … etc)
6. Flexible (multiple forms and knowledge dimensions depending on user need)

Strategic Goals and Planned Course of Action / D

The semantic wave embraces four phases of web growth:

- **Web 1.0**, was about **linking information** … **(Done)**
- **Web 2.0** is about **linking people**. **(Partially Done - AI Powered Social Web)**
- **Web 3.0**, is starting now, and it is about **linking knowledge**. **(GETCA Current Main Course)**
- **Web 4.0** will come afterward … and it is about **linking intelligences** in a ubiquitous web where both users and things can reason and communicate jointly." **(GETCA Next Goal)**

Web 5.0 will come finally and it is about **(Connecting Models)** in a " **Global Understanding Environment** " **(GUN),** which will be such upbeat, self-managed evolutionary Semantic Web of Things, Users and Concepts where all kinds of entities can understand, interact, serve, develop and learn from each other. **(The Industry Ultimate Goal)**

Strategic Goals and Planned Course of Action / E

G.A.S.E.T Project

Interactive, Intelligent, and Innovative Search Engine

Important Memo.

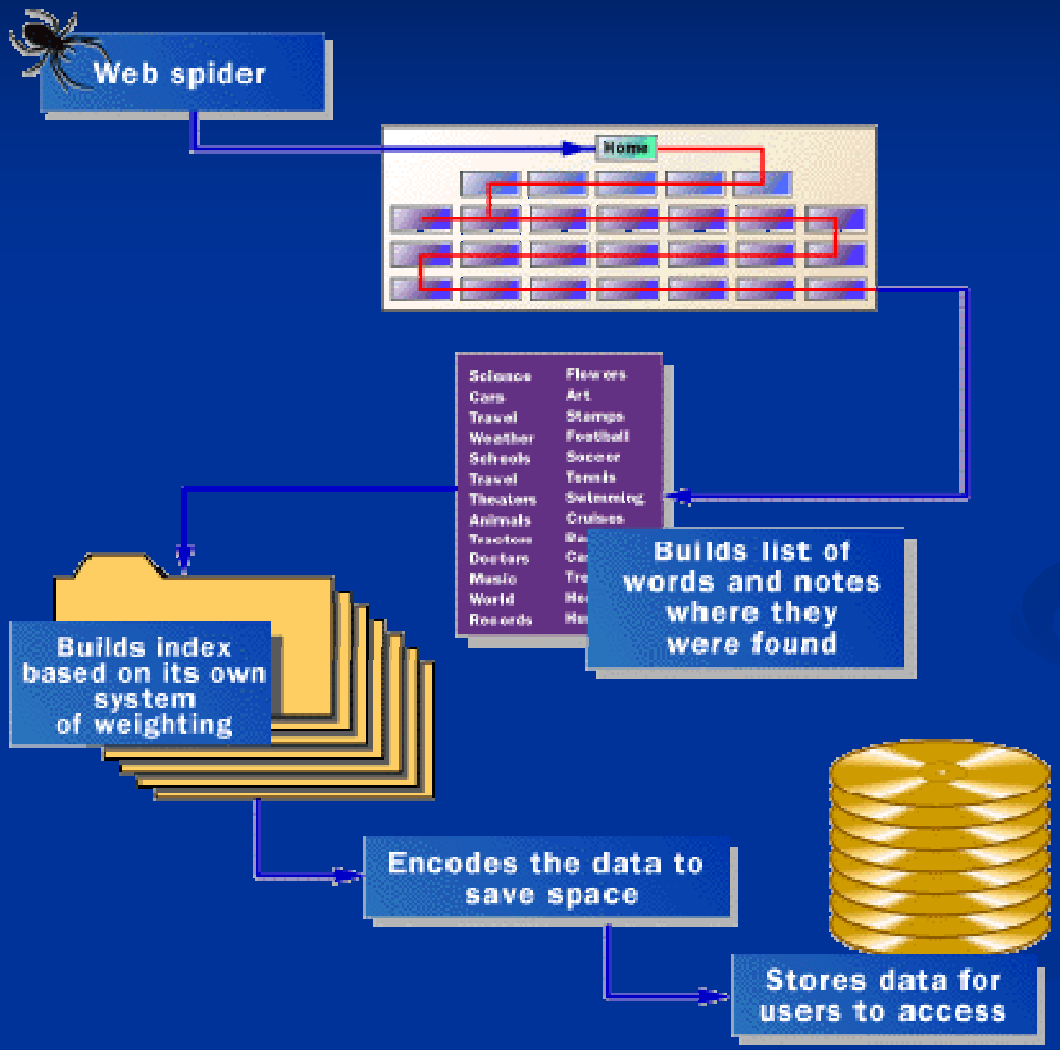
We will try to explain in non technical terms the technological differences between the implementing of (G.A.S.E.T) techniques on top of The Google™ platform- services (as G.A.S.E.T-G) compared to our **Main** strategy of having our own Servers which utilize our specially analyzed version of web pages repository.

Theoretically both are search engines, yet the need to Exclude Google ™ Results Ranking factors **"Influence"** on the received query results, which might be good for the keywords based Search engine, yet not so good for Search Engines with **(AI Powered)** Capabilities, and the need to "Redo" Major search engines tasks like: Crawling, Parsing, Analyzing, Indexing … etc) by our **G.A.S.E.T-G** system, is a (Time Consuming) task and it get in conflict with some of our main functions.

Still we had to prove that we have a functional search engine foundation, with its main Tools not only in Theory but with Indisputably Functional Applications (C# and a Intranet-Web Based Java versions), to reach our **Fully Functional** System all what we have to do is to **"Fine-Tune"** our equations, parameters and other major factors (semantic, contextual ,etc) in accordance with our project guidelines.

Traditional Search Engines Building Blocks

Main Components / Functions Diagram



In order to present the value of our Technology/product, we will have To show Standard Search Engines Main functions and technologies and compare it to ours

- Basic Keywords Based Search
- No Question Answering Tech
- No Interactive Capabilities

Traditional Search Engines theory and search techniques - 1

- Finding valuable information on the -World Wide Web is something a lot of of us take for granted. According to the Internet research firms, there are virtually 168,000,000 active Web sites on the Internet today. The job of sifting through all those sites to discover helpful information is enormous. That's why search engines use complex algorithms -- statistical instructions that tell computers how to do assigned tasks.

- Google's algorithm does the work by searching out Web pages that have the keywords the users used to search, then assigning a rank to each page based on several factors, including how many times the keywords appear on the webpage. Higher ranked pages appear further up in Google's search engine results page (SERP), meaning that the better links relating to your search query are hypothetically the first ones Google lists.

- Google's keyword search task is comparable to other search engines. Automated programs called crawlers check the Web, moving from link to link and constructing up an index page that includes certain keywords. Google then references this index when a user enters a search query. Standard search engine lists the pages that have the identical keywords that were in the user's original search terms. Google's crawlers may also have some additional advanced functions, such as being able to settle on the difference between Web pages with real content and redirect sites -- pages that exist only to redirect the received traffic by the main server to a another targeted Web page.

Traditional Search Engines theory and search techniques - 2

The Google algorithm's most important feature is arguably the PageRank system, a patented automated process that determines where each search result appears on search engine return page. Most users tend to concentrate on the first few search results, so getting a spot at the top of the list usually means more user traffic. So how does Google determine search results standings? Many people have taken a stab at figuring out the exact formula, but Google keeps the official algorithm a secret. What we do know is this:

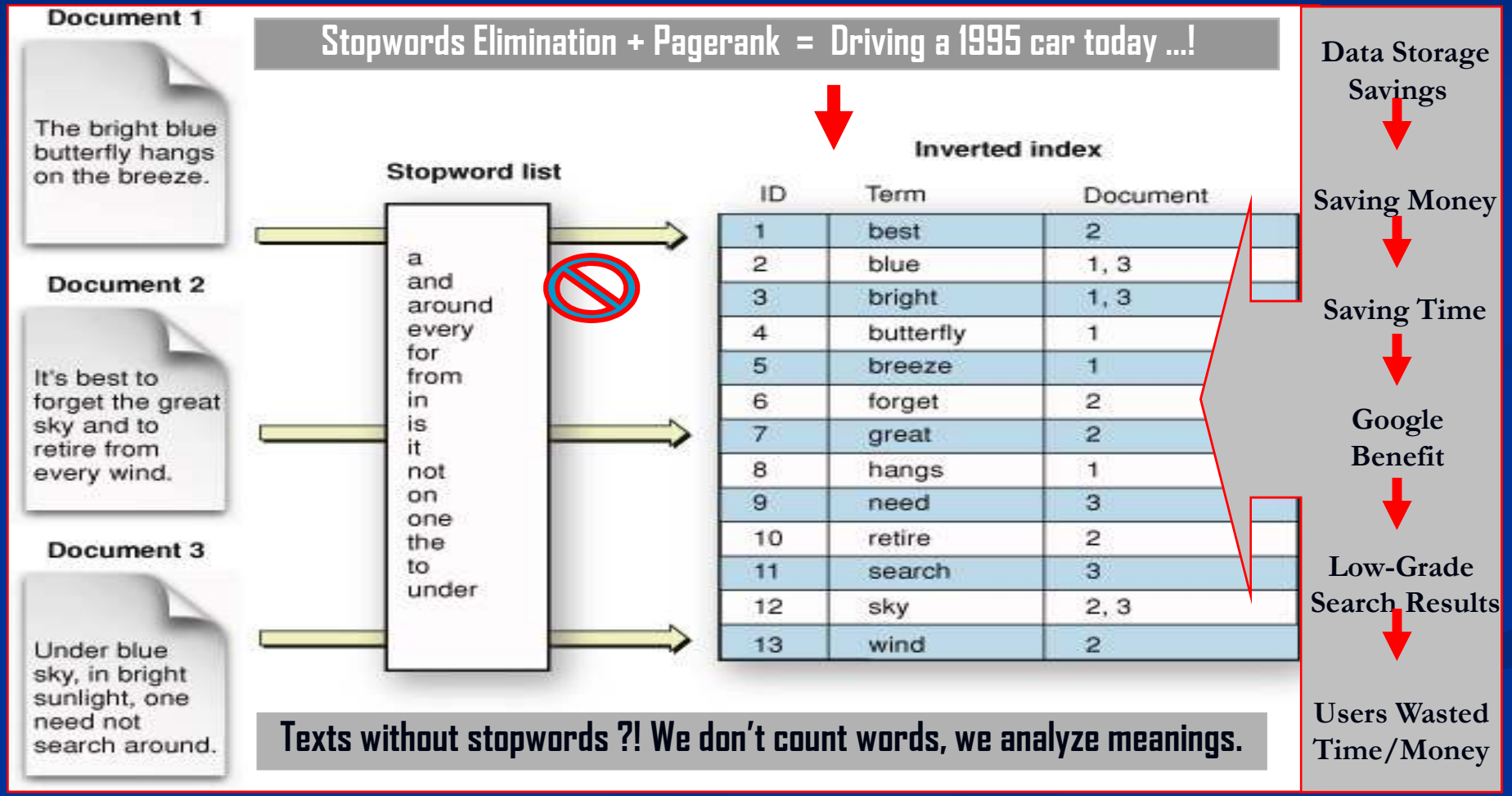
- PageRank assigns a rank / score to each search result. The higher the page's score, the further up the search results list it will appear.
- Scores are partly determined by the amount of other Web pages that link to the target page. Every link is counted as a vote for the target page. The logic behind this is that pages with high quality content will be linked to more often than mediocre pages.
- Not all votes are equal. Votes from a high-ranking Web page count much more than votes from low-ranking sites. You can't really improve one Web page's rank by making a group of empty Web sites linking back to the target web page.

Traditional Search Engines theory and search techniques - 3

- The more links a Web page sends out, the more weak its voting power becomes. In other words, if a high-ranking page links to hundreds of other pages, each vote won't count as much as it would if the page only linked to a not so many sites.
- Other factors that might affect scoring include the how long the site has been around, the power of the domain name, how and where the keywords show on the site and the age of the links going to and from the site. Google tends to place more value on sites that have been around for a long time.
- Keyword position plays a part in how Google finds sites. Google looks for keywords all over each Web page, but several sections are more important than others. Including the keyword in the Web page's title is an excellent idea, for example. Google also searches for keywords in headings. Headings come in a range of sizes, and keywords in larger headings are more precious than if they are in smaller headings. Keyword dispersal is also important. Webmasters should avoid overusing keywords, but several people recommend using them regularly throughout a page.

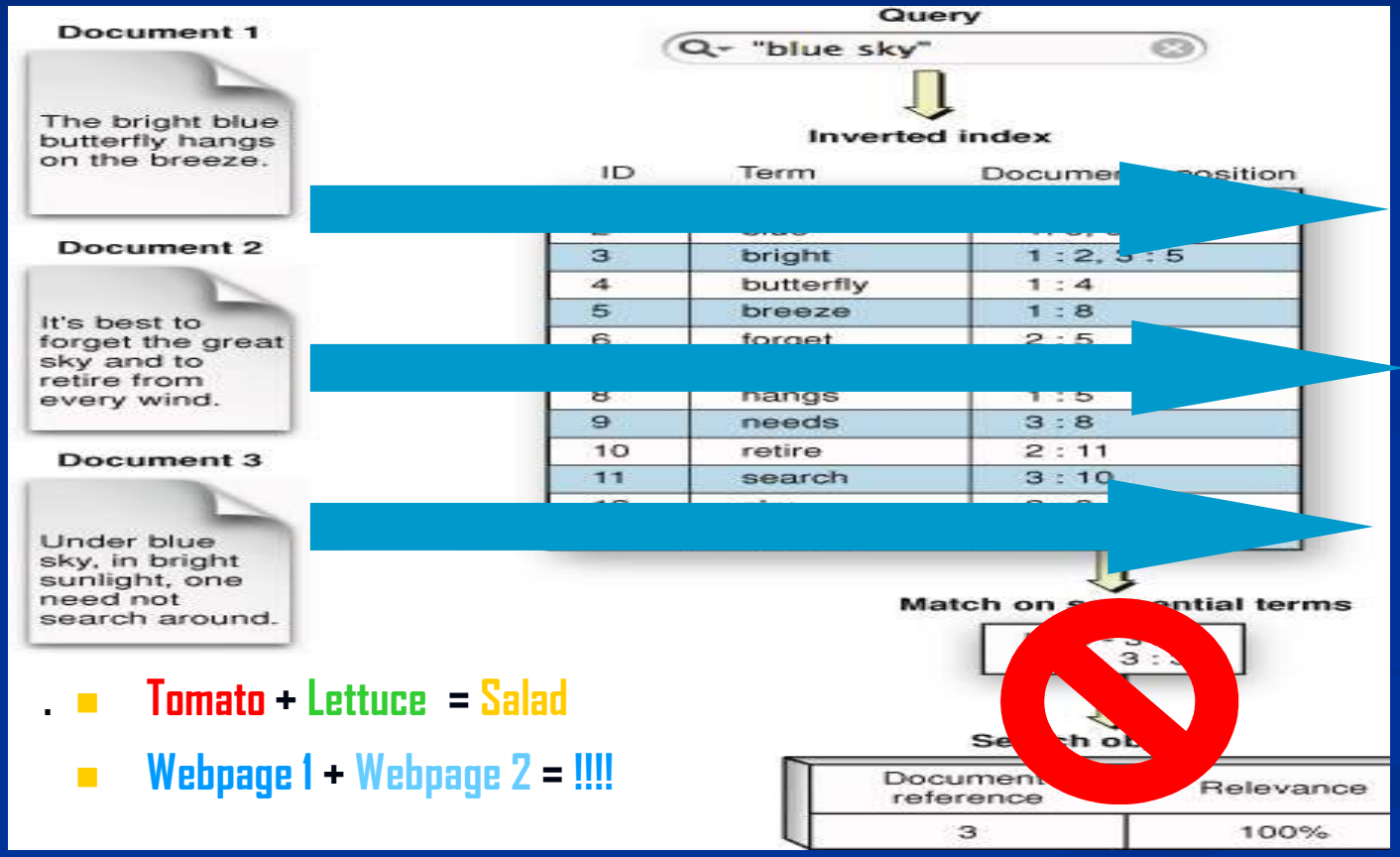
Under the Hood of Standard Search Engines (1)

Creation of Inverted Index = Destruction of Web Page Context / Concept.



Under the hood of standard search engines (2)

The indexing scheme … where is the beef !



- Tomato + Lettuce = Salad
- Webpage 1 + Webpage 2 = !!!!

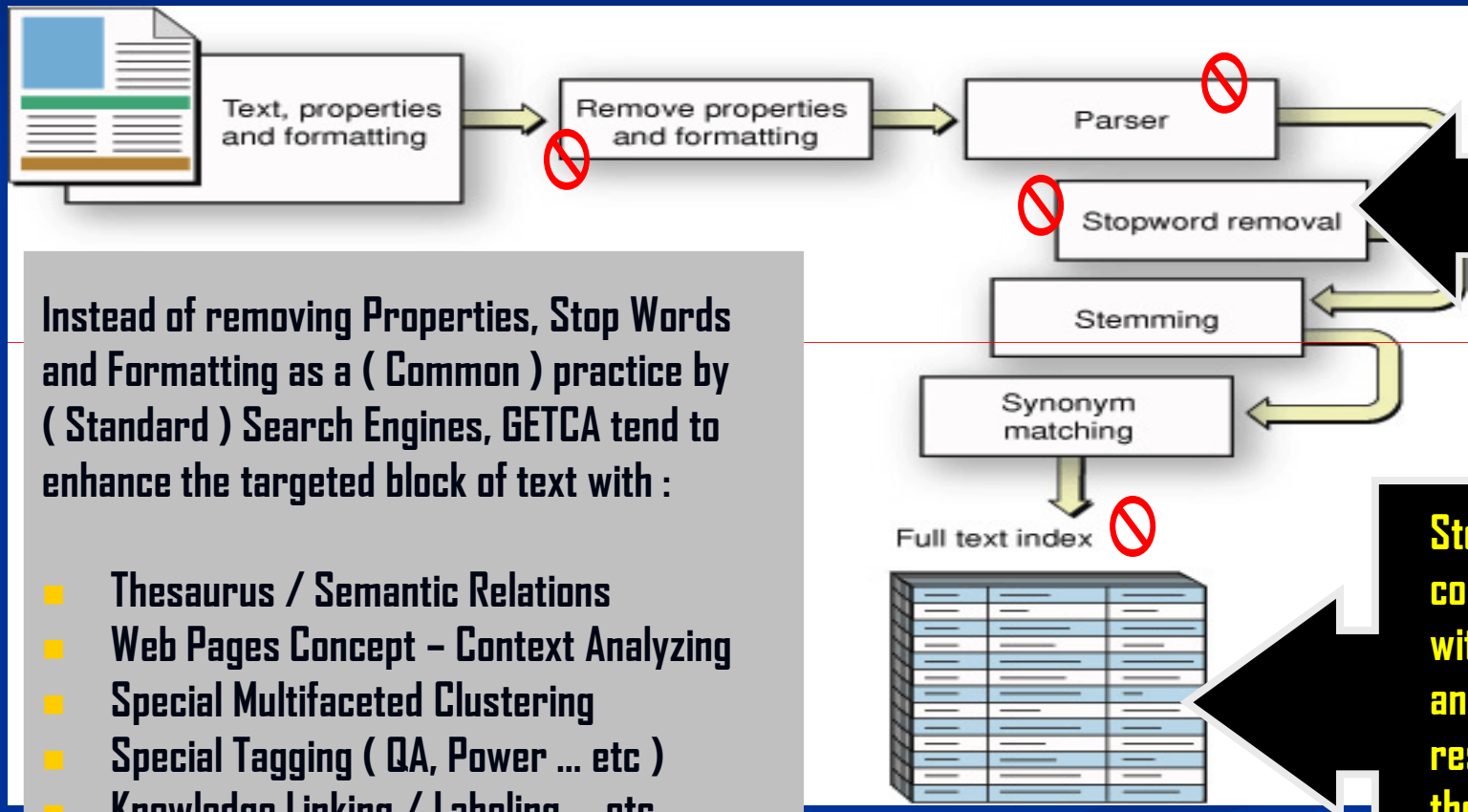
GASET Unique Tagging:

- Texts Semantic
- Concepts Linking
- Context Relations
- Finding Ontologies
- Dynamic Ranking
- Interactive Tech
- Dialogue Based
- Powerful AI - NLU
- Comprehensive QA
- Notice Multifaceted
- Apply Multitasked
- Multimedia Files

Mixing documents together will destroy web pages unique semantic leading to week clustering techniques

Under the hood of standard search engines (3)

Quantity or quality ? … do we have to choose !



Instead of removing Properties, Stop Words and Formatting as a (Common) practice by (Standard) Search Engines, GETCA tend to enhance the targeted block of text with :

- Thesaurus / Semantic Relations
- Web Pages Concept – Context Analyzing
- Special Multifaceted Clustering
- Special Tagging (QA, Power … etc)
- Knowledge Linking / Labeling … etc.

The G.A.S.E.T Approach: Multi Faceted Parsing = Multi Faceted Concepts

Storing massive collections of texts without the proper analyzing will give results similar to the one we receive from phone books.

G.A.S.E.T Project

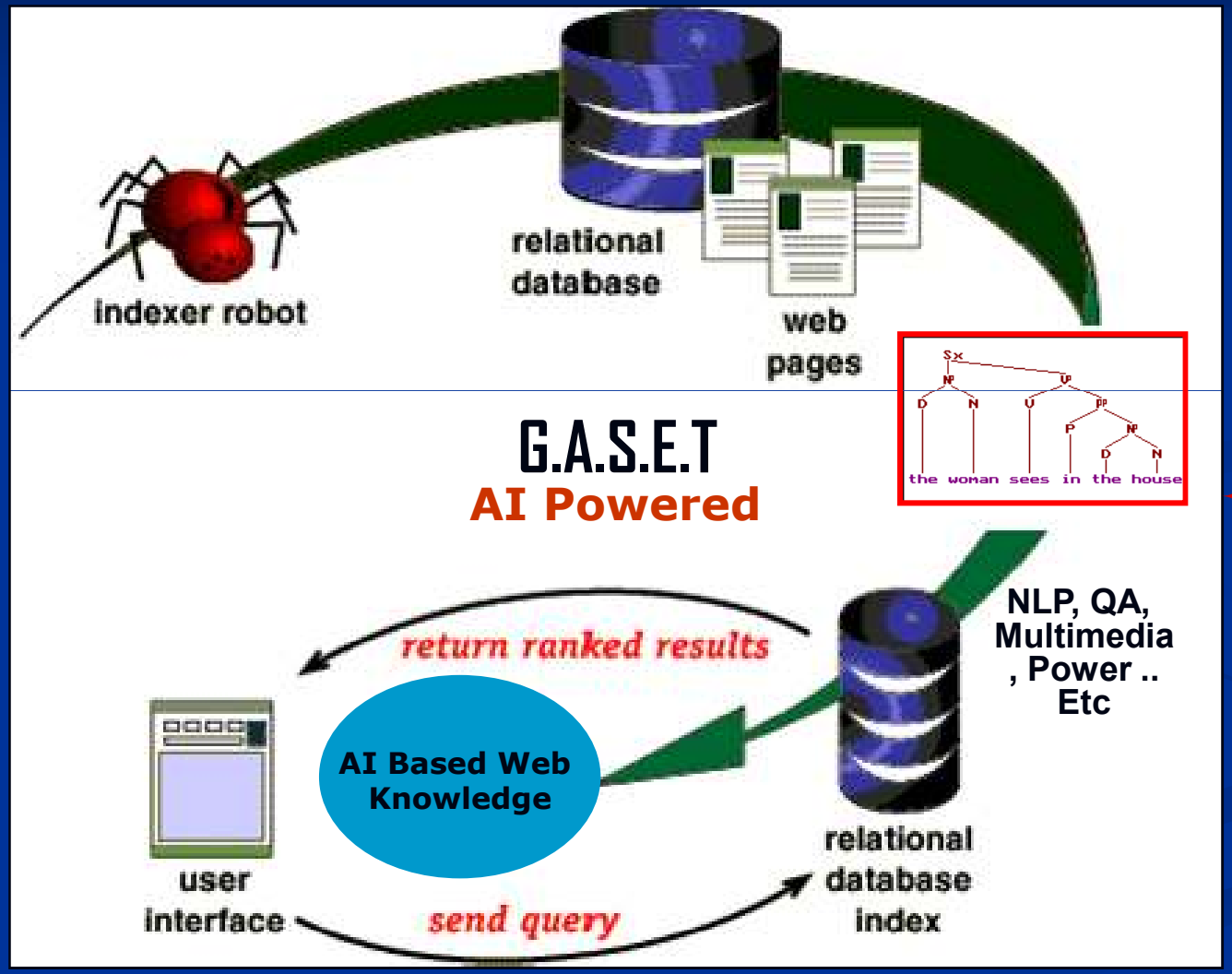
(Main Technologies – Services)

G.A.S.E.T Main System Components " Detailed Diagrams "

- 1- Detailed System Components List .
- 2- Detailed Operation Functions.

This is an attempt to describe the Complexity of our project in a Non technical terms .

G.A.S.E.T Main Components / Functions Diagram



The G.A.S.E.T way :

Concepts (not only keywords tagging)

Special parsing and crawling rules and configurations.

Auto-modified and multifaceted web knowledge base

Dynamic concept, context and sense extracting methods.

Text Retrieval and Mining (Not Implemented by Standard Search Engines)

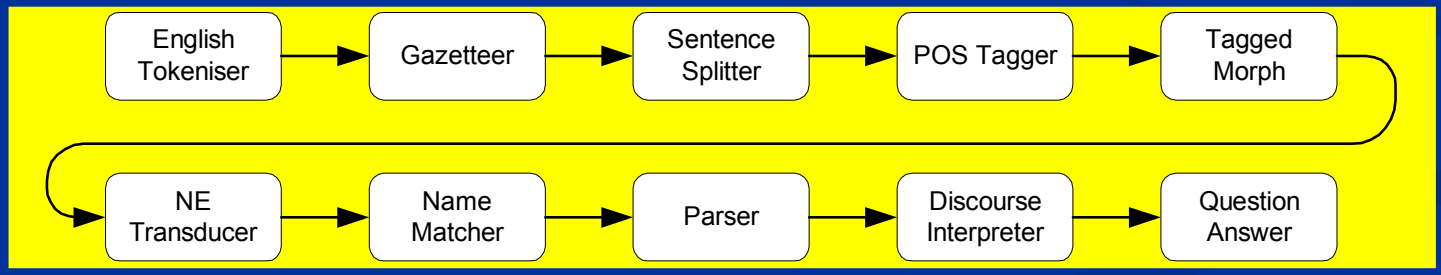
Textbooks Methods Will Lead to Conventional Results

- Lexicon terms
- N-Grams
- Pattern
- Mapping
- Matching
- Syntactic

The Solution ...

An Interactive AI Based Technology With Autonomous, Dynamic and Semi-Supervised Learning Capabilities.

- Boolean
- Probabilistic
- Vector Space
- PageRank
- Language Models



How we applied our "AI Based" NLP – NLU goals (1)

(General Description)

We designed our language technology to add value to data by:

- Value filtering
- Augmentation (providing metadata)
- Interpretation
- Transformation

We took in consideration :

- Information integration needs NLP methods for coping with ambiguity and context while taking in consideration The amount of information in textual form
- Wrapper induction is usually regular relations which can be expressed by the structure of the document
- Sites are numerous, and their surface structure mutates rapidly .
- Information that *could* be represented in a structured semantically clear format are rare

How we applied our "AI Based" NLP – NLU goals (2)

Effectiveness [Fulfilling the user needs]

Relevance and accuracy of results

- Stemming
- Spelling correction
- Query expansion - overcoming information deficit & increasing the scope of relevant information:
  - Query classification
  - Automatic thesauri
- Contextual indexing (LSI)
- User profiles (ML)
- User feedback on effective retrievals

+

- We will literally read - analyze the web for you.
- We will not give you 10000000000 answers in 0.00000000001 second, we will give you an answer because you query is a question after all .
- Ranking depend in the soul of your query not the size of the words and it counts, we should be careful about the page rank hoax … after all knowledge is commonsense not only popularity .

How we applied our "AI Based" NLP – NLU goals (3)

Characteristics of Web Content Searching

- Content is created by varied associations and individuals
- Information on the Web is inherently diverse
- Content is dispersed on several servers in numerous locations, various formats and languages aimed for diverse audiences and purposes
- (In its **April 2005** survey NetCraft received responses from **62,286,451** web sites)
- The "Open Web" of billions of static Web pages is indexed and searched via many search engines and directories

- **"Broadcast" or "Federated" search:**
  - **List of results**
- **Merging and Ranking**
  - **Increased coverage**
- **Result Clustering**
  - **Focused drill-down**
  - **Dynamic Query Models**
- **Semantic and Pragmatic Intelligence**

Web Page Concept Extracting / Enhancing Techniques - 1

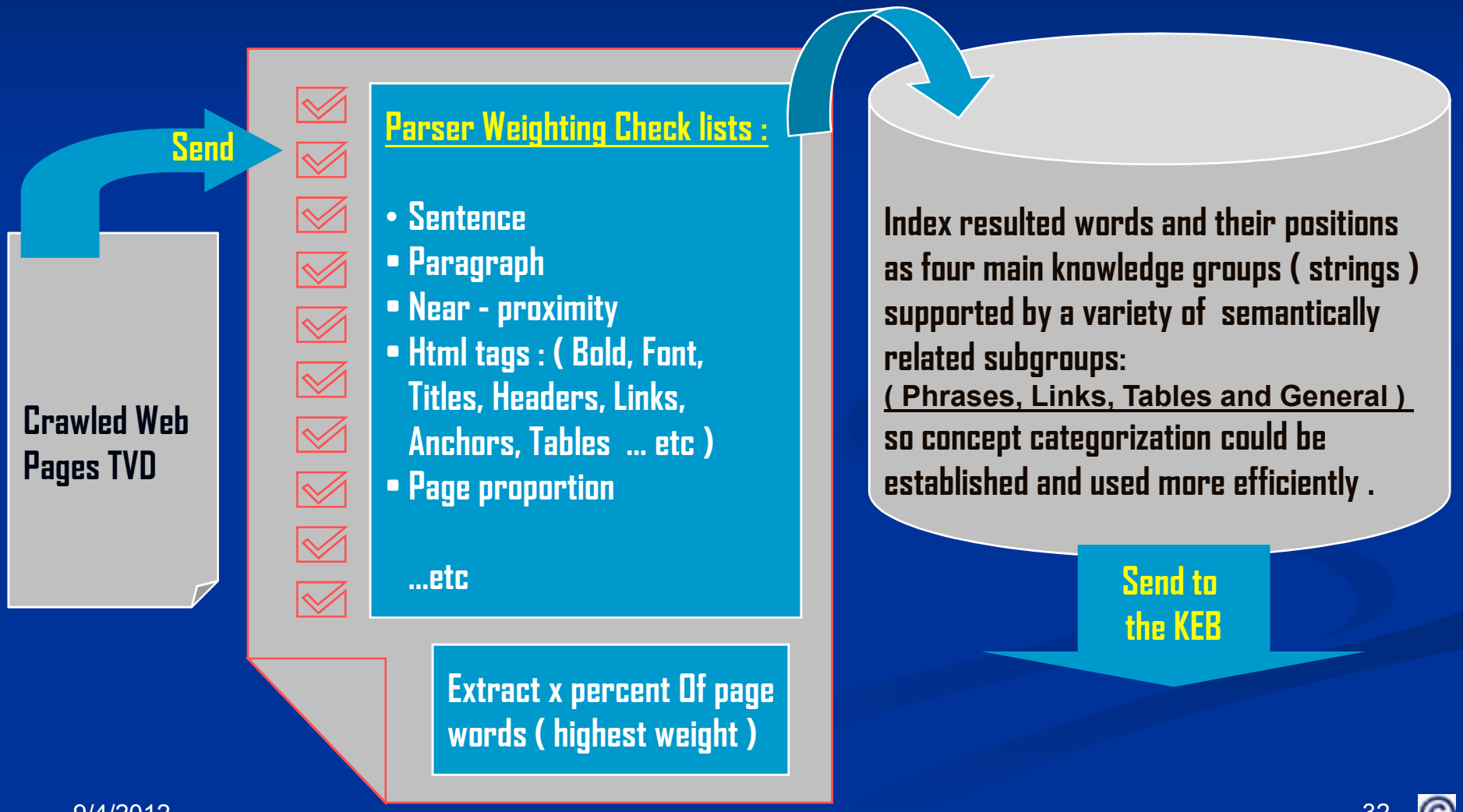
1. We use the parser to analyze the web page words and html tags, the connection between the web page parts and the related pages

2. The system will then determine grammatical-linguistic value for the page words according to their location on the page (phrase, table, links ..etc) and **relation to** each other using its specialize KEB's (Knowledge Expert Base), thesaurus and related words. It will also take into consideration the CCD (Changeable Concept Directories) and GDM (G.A.S.E.T ™ Directory Map) to **confirm** the linguistic concept and the WCDS (Web Concept Determination System) taking in consideration the KEB I values.

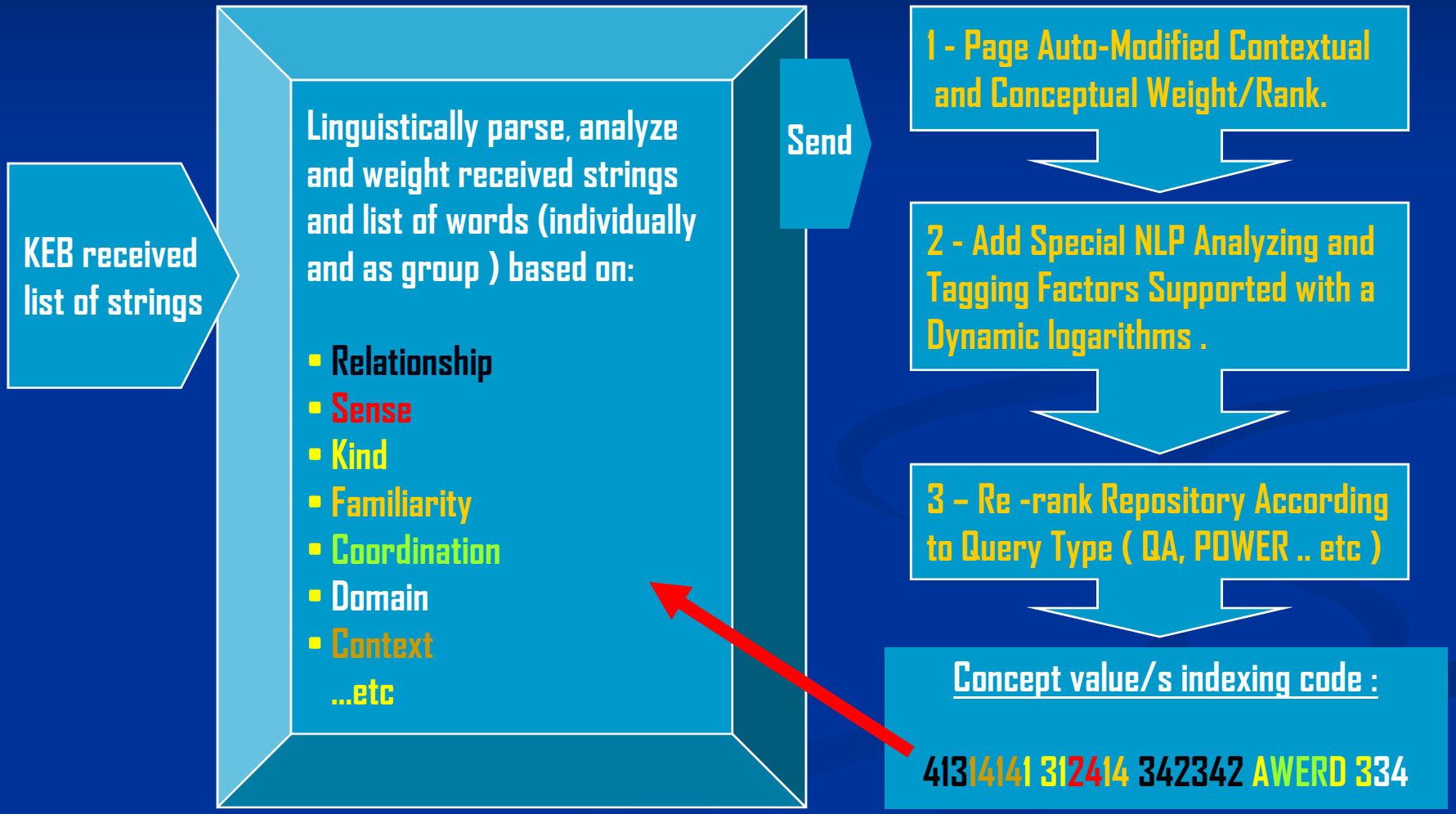
3. We will then calculate the concept of the page by receiving the value from one and two above while taking into consideration the fact that we deal with the web page **as blocks of information** with its own values

4. The received parameters and data will be passed to ASM (Advance Search Modifier) for the second query modification (to check query possibilities). Finally the system will store calculated value of concept as **XYZ numerical figure** which will determine its structure and relations to other concepts (semantically and contextually), such values will be related to ASM possible query formats.

Web Page Concept Extracting / Enhancing Techniques - 2



Web Page Concept Extracting / Enhancing Techniques - 3



Web Page Concept Extracting / Enhancing Techniques - 4

Main Functions

- QA
- Power Search
- Topic Search
- Multimedia Search

Power Search

QA

Topic Search

Multimedia Search

Web Page Tags Repository

G.A.S.E.T ™ Tagging :

- Lingo block
- Html Tags - Links
- Multimedia Forms
- Semantic
- Context ... etc

The ability to parse web pages to meet our main functions related multi indexing factors

To match web surfer possible formats of queries

Web Page Concept Extracting / Enhancing Techniques - 5

Extracted / Analyzed Concept Clusters :

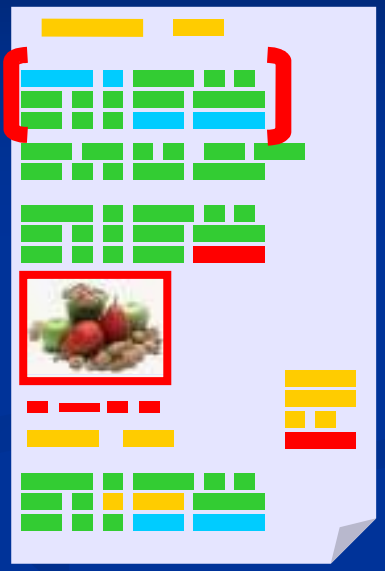
1. 42342 3ERTER422
2. T453 4534ETER53
3. 3423 4RWERW342
4. 3478 678ERW366

... etc

Concept Sorting Factors / Formats:

- Linguistic
- Context
- Structure
- HTML Tags
- Domain
- Ontology

... etc



- Verify
- Confirm
- Store

Enhance Each Concept (Context / Semantic)

(((423423ERTER422)))

- 423423ERTER422
- 423423ERTER422
- 423423ERTER422

Concept Maps

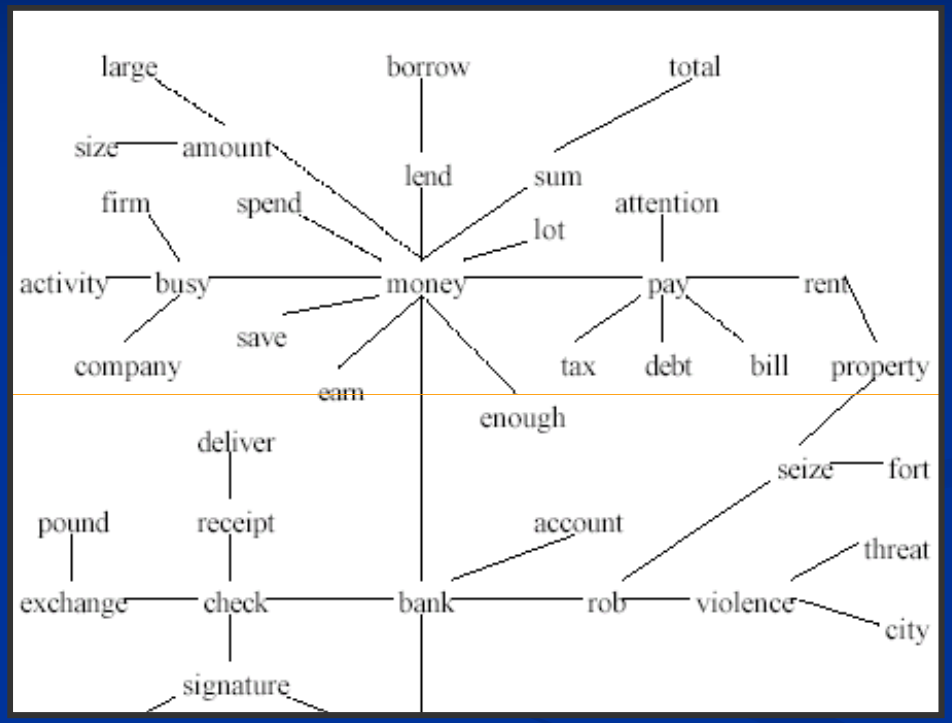
Web Page Concept Extracting / Enhancing Techniques - 6

Main Functions :

- Enhance
- Confirm
- Expand

... etc

Compare to the Main Repository of Concepts Barrels



Concept cluster # 1

- 23etr5645666

Concept cluster # 2

- 6f5dsdf4e48d4

... etc

Compare and Match User Query to Result

The Technology

- A - THE PROCESS .
- B - DOCUMENTATIONS .
- C - CODE INFORMATION .
- D – THE TECHNIQUE .

A- The Process

The main goals behind our project are to: expand the web surfer linguistic / context query options, and to enhance his / her web searching frontiers by applying the following techniques:

- Improve and enhance the query's entry by creating pull down / scroll tool bars that incorporate lingo suggestions (thesaurus, related words and directory concept maps).
- The system then stores and analyzes the crawled web repository according to our natural language processing (NLP) parameters, specialized contextual factors and advanced web tagging techniques (using indexing framework and customized concept based knowledge expert bases) which consider the logical factors of web pages analyzing.
- The results will be presented back to the user in what we think is improved way of ranking, and he / she will benefit from using our intelligent and comprehensive searching tools.

B-Documentation

- As you will see in technical documents supplied later on, (GET-Jo) has verified all the potential problems facing the Google ™ intelligent information retrieval and worked on finding solutions that took into consideration the complexity and diversity of the web in a way no other company has been successful in achieving.

- We have documented our project in a paper, which is being used as an comprehensive guideline in explaining our project. The paper has been registered with the official Jordanian patent organization . Future international patenting is expected before the end of 2009.

C - Code Information / 1

- **Two versions of G.A.S.E.T ™ project where made :**

1. C# desktop copy with minimum hardware requirements (for main functions evaluation and testing purposes - which doesn't require a web server).
2. Java web based copy with limited web repository (indexing limited to server memory capacity and more active functions (to test the actual testing of the project as a web server)

- In regard to programming, we built our systems (C# and Java) from **scratch** (**Crawler, Parser, Analyzer, Page Ranker, lingo verifier … etc**), one exception was with the indexing whereby we are temporarily using the LUCENE open source / apache licensed platform with major code improvements by our programmers .

- We also incorporated multiple **modified and dynamic** linguistically / conceptually related dictionaries and directories in our Java language web based project, its interchangeable design and flexible structure was integrated smoothly in our code as a major support for our analyzing engine.

- Pioneer technologies developed by our programmers (html table parsing, multifaceted indexing … etc) where documented in the code core in away that it could be understood easily by other programmers.

C - Code Information / 2

- It is worthy to note that our system takes in great consideration the html tags provided in the web page, weigh them in regard to their importance, context and comparable page value, we also provided a solution to the html table analyzing and extraction difficult task through the latest clustering / parsing techniques. By doing so, we were successful in improving the received results. We also used our own threading /speed optimizing techniques to cut the processing time .

- The use of multiple developments packages, in order to **speed** the coding design, generation and testing was of great help to us in our quest to create a projects that meets the highest programming standards of quality.

- **Rigorous testing** where executed (taking into consideration the limited hardware / web storage capacities we are currently facing), tests where preformed by our testing department and regular web surfer who confirmed for us the great results we expected from the currently implemented technologies and the superior results we received from our tasks simulators group who preformed stressful function completion.

D – Technique / 1

- We designed our own versions of Html Crawler and Parser to guarantee search precision and compatibility. We understand that the web page knowledge is based on *discarded pieces of information. We also understand that t*he task of knowing the web page or site identity is challenging process. Therefore, the precise and open minded approach in separating important information from the rest of the data was essential in developing our system.

- We also took the step of analyzing competitors system's source codes and documentations (Nutch, Red-Piranha, Websphinx, JWNL ..etc). We took their points of strength into consideration and avoided their disadvantages. Our engineers were successful in utilizing the best available "Data Mining Technology" available while keeping the simplicity aspect of Google ™ 's Interface intact.

- We have also devolved our system to recognize the questions form of queries (QA) in a compatible way with the web construction, keeping the results speed retrieval advantage intact.

- We have adopted the complex NLP Analyzer System to aid us in analyzing, by using phrasing and indexing methods, the received search results from the user's Google ™ search. Doing so enabled us to rearrange the results to represent what we think is a logical page rank.

D – Technique / 2

- We believe that that Google™ definitions of user query should be expanded in the case of verbs, adjectives, adverbs and common nouns. Such expansion will help the user in his/ her quest for more accurate results. We take pride in saying that our system was the first to incorporate such technology.

  Example: If the user entered "Build A Car", the general search engines, like Google™ and Yahoo, will most likely retrieve pages containing the two words "Build and "Car". Without concerning themselves with analyzing of **the concept of the query.** This might get us some good results but it could also give us a poem about Building a car or toy car builders. Our approach, on the other hand, will be seeking pages and websites that discuss the building of cars as a concept, entities specializing in car building…etc.

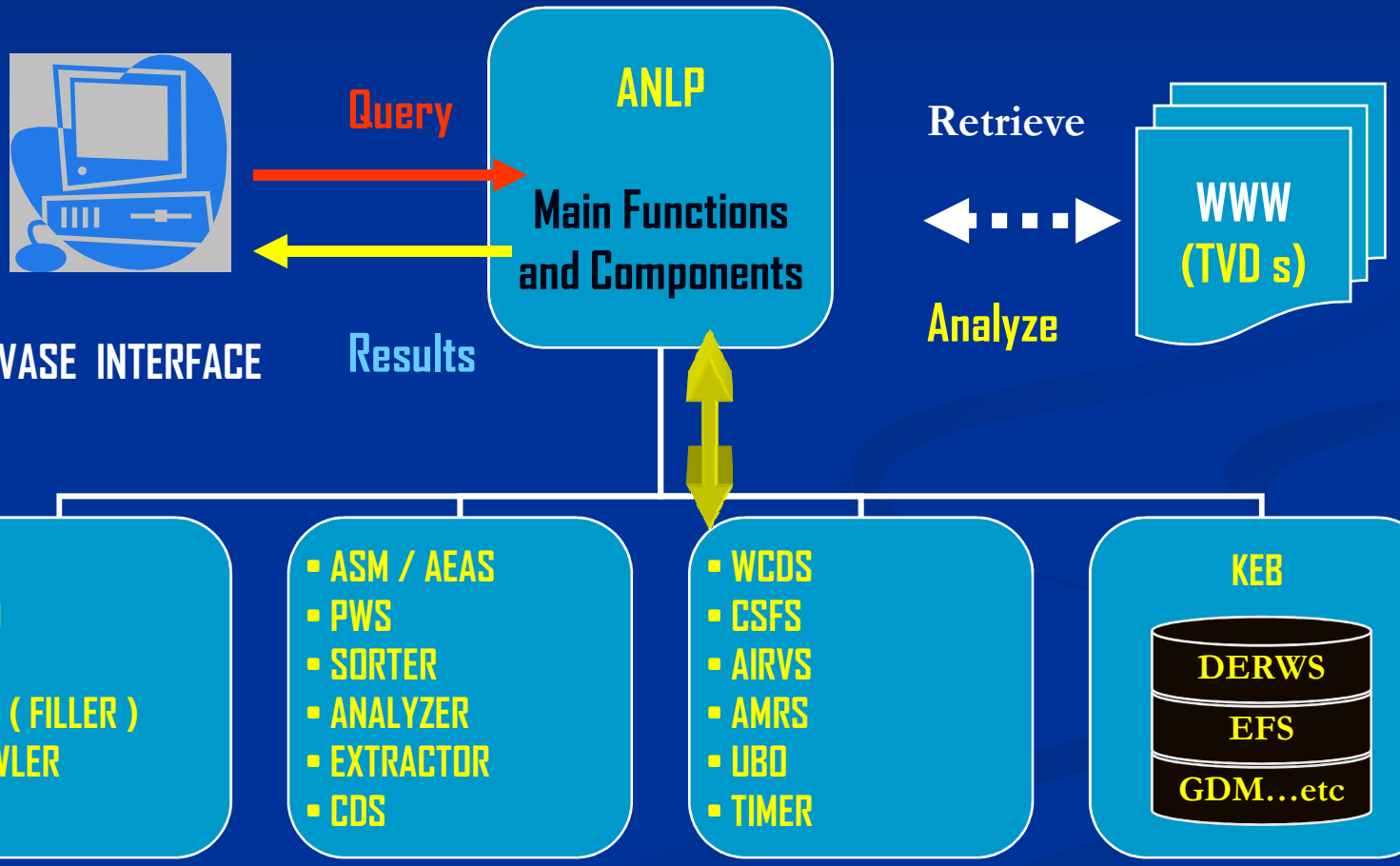
- We were successful developing a system that takes the query concept into consideration while paying attention to **Google 's ranking**. After all, Google ™ and the other main search engines are the base of our system's base.

- We also have designed our project in a away as to be adaptable while adding the benefit of using our system's suggestions and unique system of multiple queries and analyzing to improves query results.

G.A.S.E.T Project

(Main Components)

Main Components / Functions Diagram



Examples of Intelligently Crawled Corpora

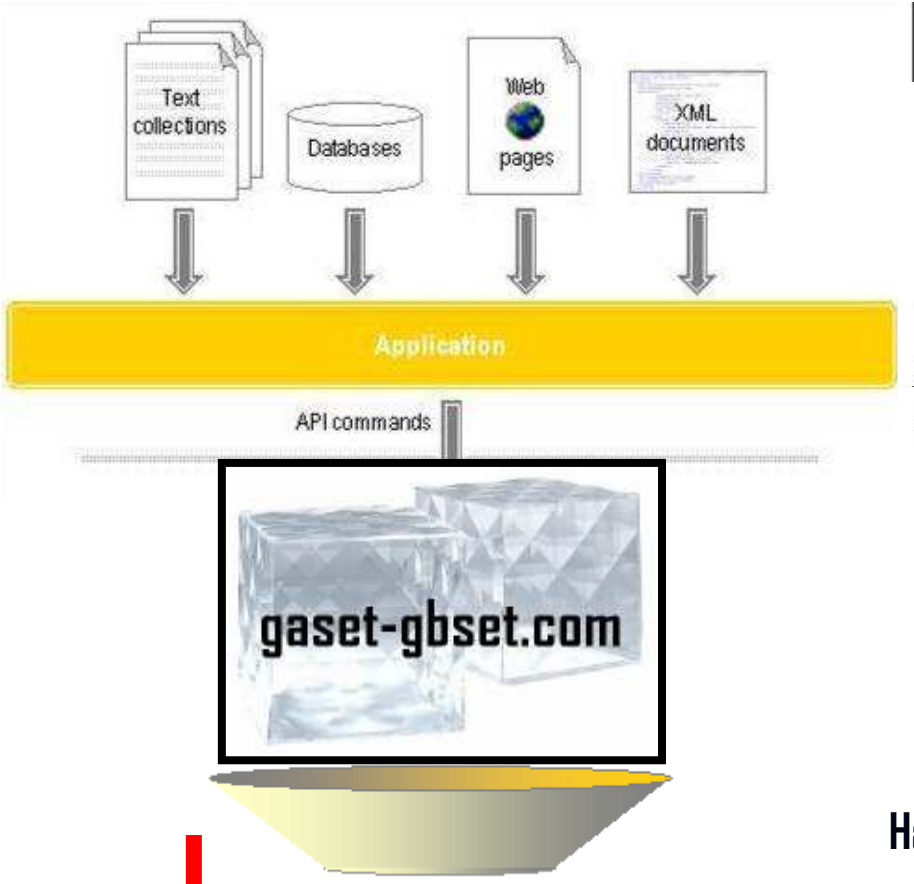
- Netflix challenge
- AOL query logs
- Blogs
- Bio papers
- AAN
- Email
- Generifs
- Web pages
- Political science corpus
- VAST
- del.icio.us
- SMS
- News data: aquaint, tdt, nantc, reuters, setimes.
- Europarl multilingual
- US congressional data
- DMOZ

- Timebank
- Wikipedia
- wt2g/wt10g/wt100g
- dotgov
- RTE
- Paraphrases
- GENIA
- Generifs
- Hansards
- IMDB
- MTA/MTC
- nie
- cnnsumm
- Poliblog
- Sentiment
- xml
- epinions
- Enron

The web is really large :

- 100 B pages / New pages get added all the time
- Dynamically generated content
- The size of the blogosphere doubles every 6 months
- Yahoo deals with 12TB of data per day (according to Ron Brachman)

Software Main Components Interaction



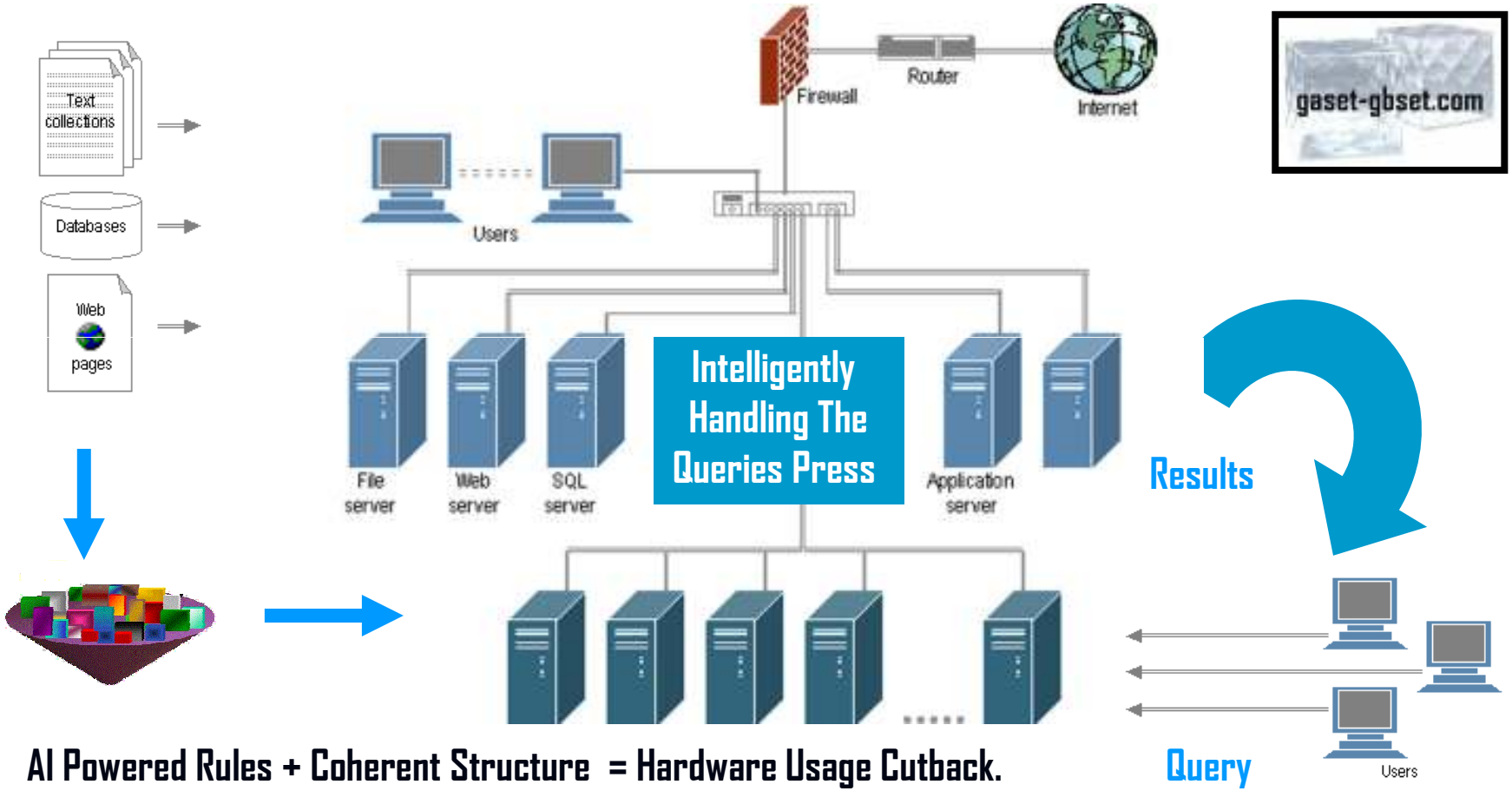
Webbase Knowledge Harnessing, Enhancing - Dynamic Utilization

AI Based " Ranking Rules " / Concepts Creating - Modifying

Building Repositories to be Utilized by our Dialogued Based "Interactive" Tech

Hardware and Software Interaction

Hardware Main Components Interaction



AI Powered Rules + Coherent Structure = Hardware Usage Cutback.

Main Components - 1

- **VASE:** Visual Advance Search Enhancement.
- **KEB:** Knowledge Expert Base

Which consist from the following parts :

1. **GDM:** G.A.S.E.T / DMOZ ™ Directory Map (to determine concept)
2. **DERWS:** Dictionary/Encyclopedia Related Word System (auto-modified sources)
3. **EFS/STOP WORDS:** Exclusion Function System
4. **PHRASE KEB:** (logically combined and interchangeable database)

- **TVD:** Temporary Virtual Directory. (specialized collections of web URLs)

Main Components - 2

- **ANLP :** Advance Natural Language Processor

Which contain the following parts :

- **CCD:** Changeable Concept Directories (checks concepts group relations)
- **TCCD:** Temporary Changeable Concept Directories
- **LPGS:** Logical Phrase/near Generator System . (generate logical combinations)
- **VASE(FILLER)** (rearrange and reorganize query positions and other functions)
- **Crawler** (crawl received web pages)
- **RLS:** Re-Learning System (adjust concept by learning from received results)

Main Components - 3

ANLP continue

- **ASM**: Advance Search Modifier (modify search through logical combinations)
- **PWS**: Page Weighting System (add all the conceptual and linguistics factors)
- **Sorter:** (distribute and rearrange results and query terms)
- **Analyzer:** (analyze received results and query terms through advance equations)
- **Extractor:** (customizable and intelligent extracting mechanism)
- **CDS**: Concept Determination System. (from linguistically organized databases)
- **WCDS**: Web Concept Determination System. (from web and its directories)

Main Components - 4

ANLP continue

- **CSFS:** Concept Similarity Finding System. (multiple interchangeable sources)
- **AIRVS:** Advance Initial Result Viewing System. (showing results in multiple forms)
- **AMRS:** Advance Multiple Ranking System. (linguistics and conceptual sources)
- **UBO:** User Behavior Observation. (observe users search strategy and techniques)
- **TIMER:** (to be used with multiple operations)
- **ELIMINATE FUNCTION:** (to extract unrelated terms and stop words)
- **AEAS:** Abstract Extractor - Analyzer System. (unified and specialized techniques)

VASE Components and Functions - 1

Interacting with the user

- We start by modifying the Google ™ advance interface by building the VASE (Visual Advance Search Enhancement). The new interface will have a pull down menus that would carry the common thesaurus and related words for the user input word/s.
- The thesaurus and related words (combination from multiple sources) will be retrieved from the KEB, simple and specialized dictionary and the DERWS (Dictionary/Encyclopedia Related Word System).
- The user will have the option to search for a term in the anchor, text, title and URL and search for the rest of the terms in the document .

VASE Components and Functions - 2

Interacting with the user

VASE INTERFACE

G.A.S.E.T Suggestions

(Scroll Menus)

GDM

Connections options

or Concept Suggestions

Thesaurus

To replace original query

Related Words

Add on Suggestions

Web Related

Add on Suggestions

VASE Components and Functions - 3

Interacting with the user

- Initial concepts, will be generated by using the **CCD** (Changeable Concept Directories) ; the concepts will be extracted from the **GDM** (G.A.S.E.T Directory Map)
- We will also have the **TCCD** (Temporary Changeable Concept Directories) – Web **Related** words that will be used when the user enters a single term (similar in idea to the **GDM** – Yet generated from **G.A.S.E.T** web Repository) .
- The generated concepts will be displayed to the searcher (and then added to the interface) to aid the system in narrowing the search to the users concept choice
- Each word shown in the query will be shown in the menu as **Bold** and it will have up to four linguistic/ web extracted choices (Thesaurus, Related words and Web related), when the user chooses to modify his/her query, the system will take this into consideration and update the suggestions on the VASE interface
- The searcher's decision on the terms used in the query will have great impact on the received results, trying to find all the possible conceptual relations between the query terms and their related choices will be the main mission of the VASE function

VASE Components and Functions - 4

Interacting with the user

VASE INTERFACE
(Dynamic Scroll Menus)

KEB:

1. DERWS
2. GDM

ANLP:

1. TCCD
2. CCD

Find results

Thesaurus (replace by)	Related words (add to)	Web related words (add to)
car	car	reviews
auto	gas	auto
automobile	accelerator	feb
motorcar	wing	online
		news
		prices
		information
		buy
		rental
		search

Filler – Extractor

Related Directory Maps

Top > Arts > Movies > Titles > B > Blue_Car
Top > Arts > Movies > Titles > B > Batman_SeriesBatman_BeginsReviews >

Auto Modified Choices

System will watch the User action and it will Re-rank proposed lists Consequently, taking In consideration many Contextual - lingo Related parameters.

Query Modification - 1

The next step is to create an effective query. We built a LPGS (Logical Phrase/near Generator System) which will do the following :

1. Analyze the user original query words combination, thesaurus, related words and initial concept/s to generate the best possible combination of queries taking in consideration various factors. This is done by consulting the KEB and the ANLP.

2. Invoke G.A.S.E.T web Repository (Google ™ in the case of G.A.S.E.T - G) to find the best matching combinations taking in considerations our multiple ranking / weighting equations and parameters.

3. 1 and 2 above will need to work simultaneously to retrieve the best combinations.

4. The system will be able to re modify its own parameters and equations in line with the knowledge it acquired by supervising the user choices and actions at VASE, such knowledge will help in saving time and enhance semantic and contextual choices given to the future users.

Query Modification - 2

VASE Interface

- Thesaurus
- Related
- Ontology
- Lingo
- Senses
- Fields
- Categories

... etc

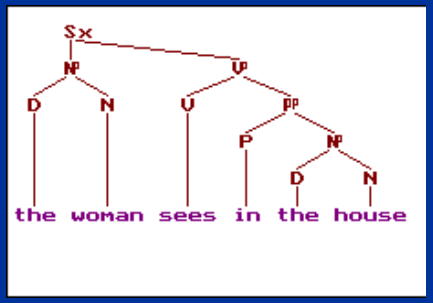
KEB:

1. LPGS
2. Analyzer

ANLP:

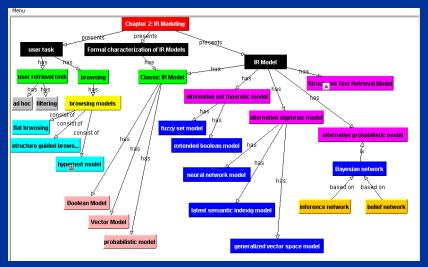
1. Concepts
2. Context
3. Contrast

... etc

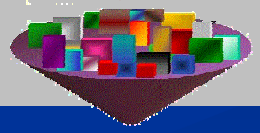


NLP Analyzer

Auto modified formulas, factors and ranking / analyzing parameters.

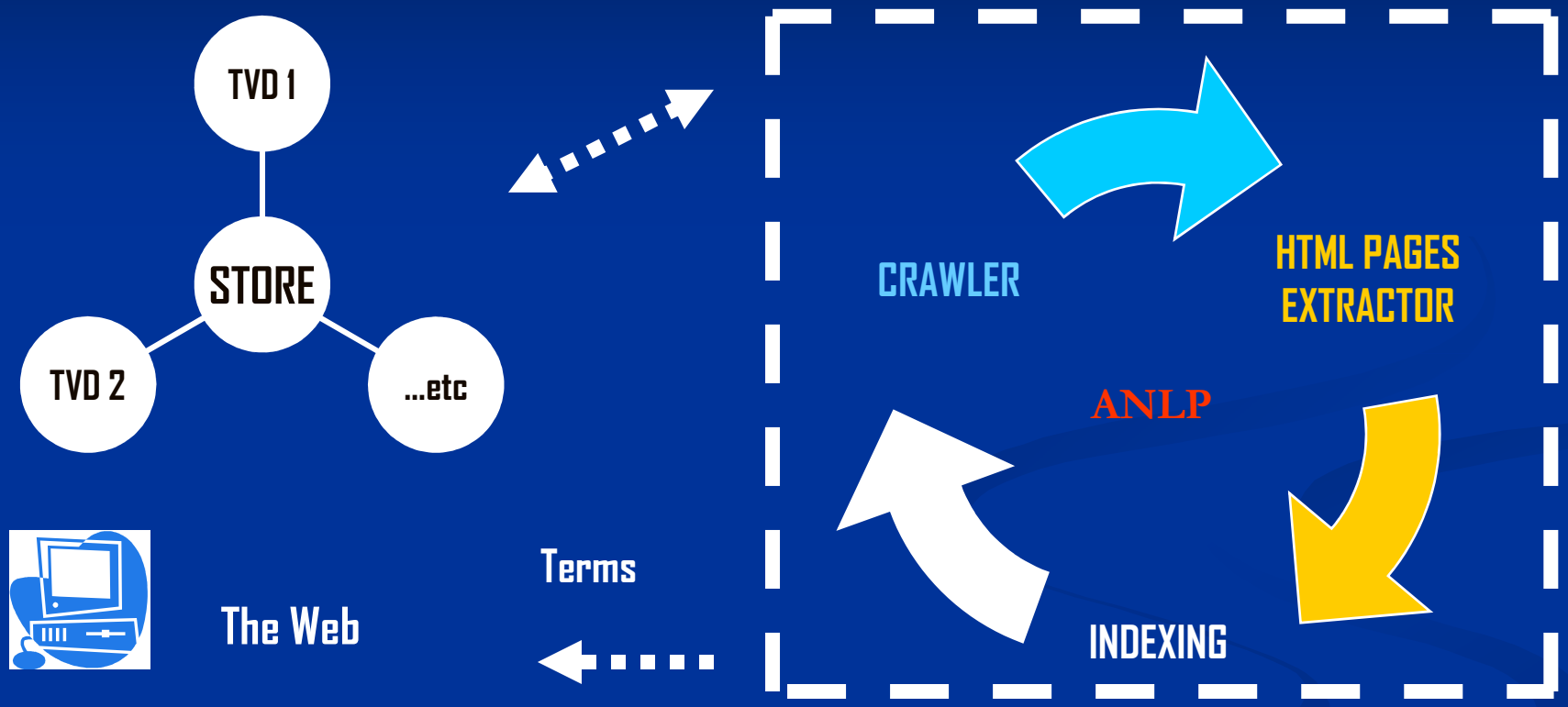


NLP Equations



Interacting with the Web Repositories

" Google ™ in the case of testing our technology by G.A.S.E.T "



Once the multiple queries are ready, the **ANLP** invokes the system to start the **Crawling Process** and the analyzer will **extract** the URL's and passes it to the **crawler** to retrieve the pages and store it in the **TVD's.**

Analyzing TVD's / First Stage - A

The major part in our project is analyzing the TVD's and the KEB to enhance the search. The ANLP starts by analyzing the TVD's in parallel by doing the following steps :

- First we apply the EFS (Exclusion Function System) that use a list of unrelated words from the KEB to classify conceptual distinctiveness TVD (in accordance with the semantic / context verifying function).

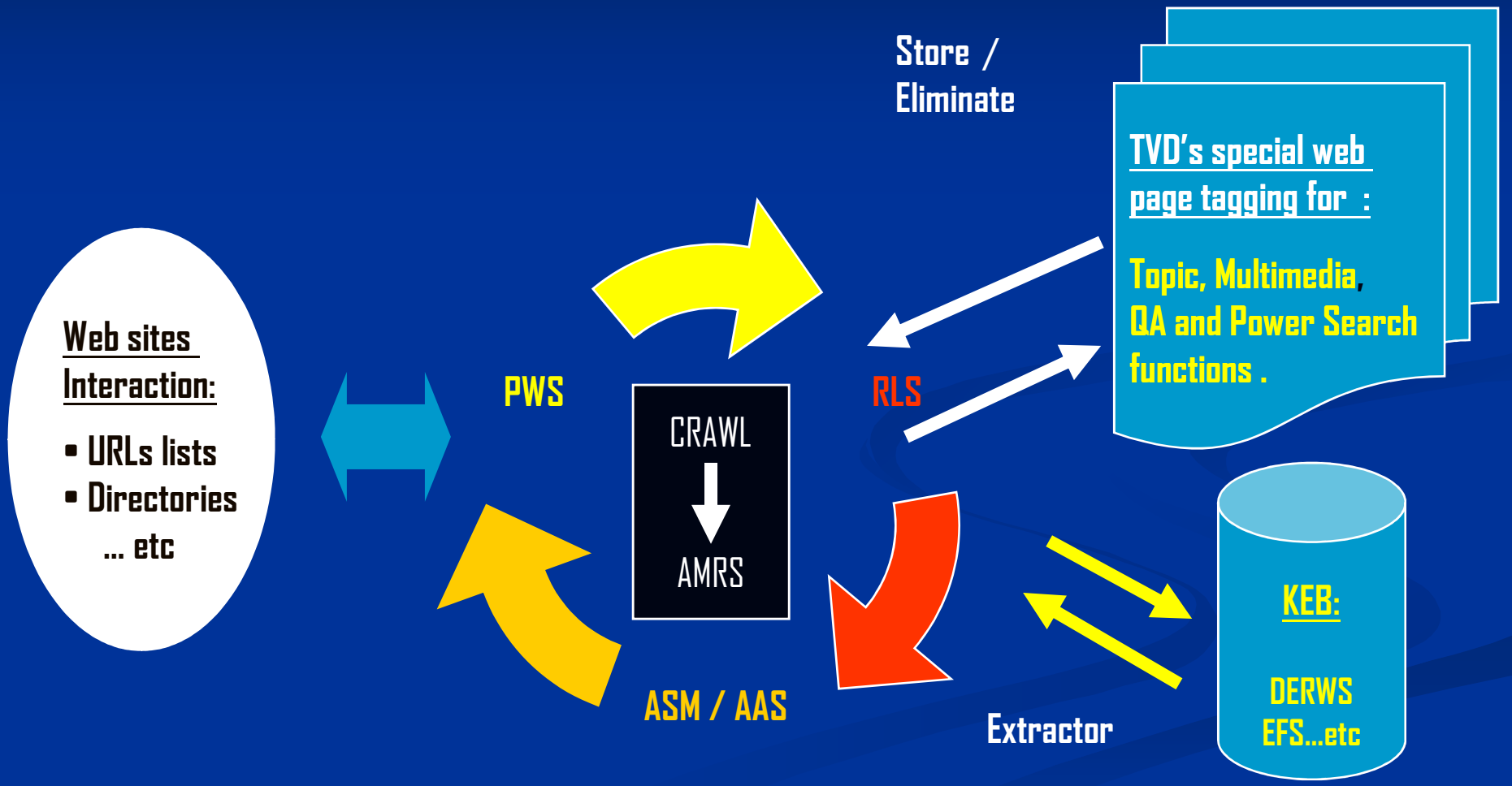
- Next we apply the RLS (Re-Learning System) that will examine the documents in the TVD's to check for special words; near the target words or in the title (the page or paragraph) to be passed to ASM (Advance Search Modifier). The ASM is a system that we need to have combined with the AEAS: (Abstract Extractor - Analyzer System) to look for some kind of pattern or contrast .

- The ASM will use the information coming from the RLS and KEB to choose possibility for new Conceptual barrels – tags . Then it repeats the interacting with G.A.S.E.T (or Google ™ in the case of G.A.S.E.T) consecutively to establish new semantic and perspective relations which meet our G.A.S.E.T main functions classifications and tagging criteria's (for the QA, Power Search ... etc),

Analyzing TVD's / First Stage - B

- The operation will be more dependable on internally generated and revised formulas, the proper design of NLP rules and functions will either make it or break it.
- The analyzer will examine the results TVD's and the **PWS** (Page Weighting System) will add page weight (as a whole and as parts) , the weight amount will be applied autonomously with the help of confirmed KEB intangible data, that will guide the essential start of the web repositories analyzing operation which will lead to the process of concept paradigm enhancement, the end results will be the fully analyzed web .
- Eventually it will pass the resulted web repositories back to the analyzer so it will invoke the **AMRS**: (Advance Multiple Ranking System) that will rank the web pages according to the needed ranking factors.
- The ANLP will start its rigorous task of building many-sided inverted index which will respond to main Functions.

Analyzing TVD's / First Stage - C



Analyzing TVD's / Second Stage - A

1. The **AIRVS** (Advance Initial Result Viewing System) will display the ranked list received from the ANLP choices of pages **while observing the user's behavior by observing his/her opened or pages not the desired.**

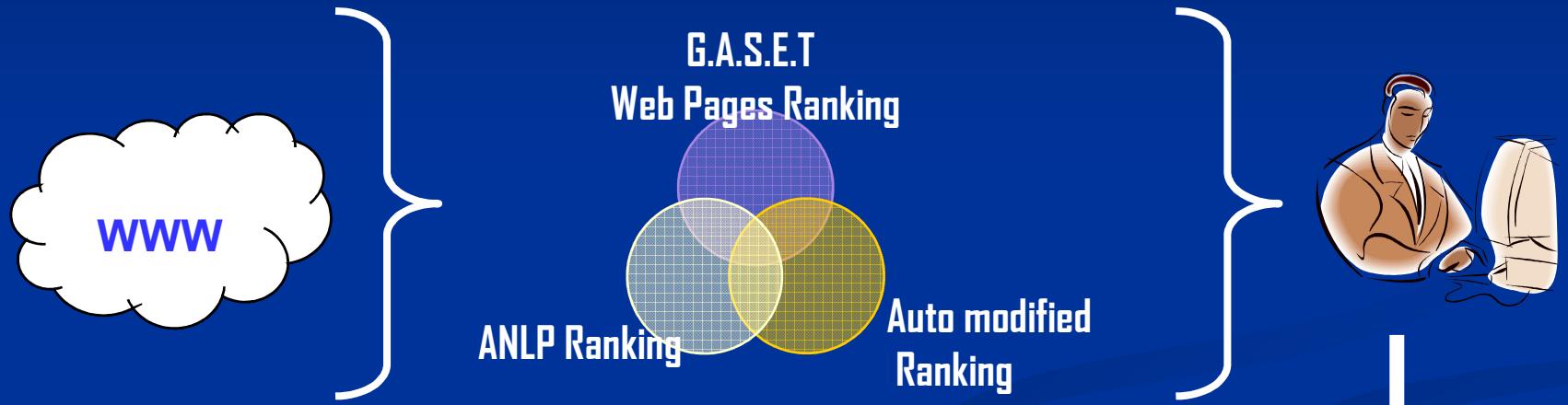
2. The **UBO** (User Behavior Observation) **will observe the time used to read the first page of result.** If the searcher chooses **not** to go to the second page before the (x-1) time limit, the UBO will consider this as probably lack of interest in the results retrieved which will lead to lowering its ranking.

3. The analyzer will initiate the UBO system, which will check which pages are opened (if the opened pages related to each other) and the time in which each page was viewed by the searcher.

4. The system will then reconsider the ranking of these pages. In the meanwhile, the analyzer will use the **CDS** (Concept Determination System) to receive -generates a concept from the pages in the TVD.

5. More patented factors will be taken in consideration .

Analyzing TVD's / Second Stage - B



The **UBO** (**User Behavior Observation**) function will be **Sensitive** to the varieties of web browsing methods and techniques used By the G.A.S.E.T users and it will be customized accordingly.

Samples of currently implemented technologies :

- Automating the whole searching process by applying **the main concept** of the search over the web through the thorough implementation of our various complicated algorithms which will enable our user to use combined G.A.S.E.T functions at the same time like … QA and Multimedia .
- Enhance our version of the indexed web repository with the needed **linguistically harmonized terms** in a verities of combinations using our highly integrated expert system, then **Re-rank** the results according to its web conceptual matching and linguistic credibility using our developed NLP techniques and parameters.
- Filtering and indexing the received web pages to **determine its concept** by analyzing the adjourned set of terms, page concept, unified information blocks (UBI) and implementing the Re-learning process and techniques to compare generated concepts .
- The web surfer **behaviour – choices** will be an essential part of our ranking, results analyzing and **logarithm auto-modifying**, it is the soul of the web .

G.A.S.E.T Enhancement for Google ™ " Similar Pages Function " - 1

When the searcher chooses to search for similar pages that interest him/her, the CSFS (Concept Similarity Finding System) will invoke the following functions:

1. Each web page (parts or as a whole) from the beginning will have its own multifaceted tagging – classification identity, such idea/tag will be dynamically designed with ability to be used in its base form (general concept) or in its enhanced structure (joined with other concepts, harmonized replacement of thesaurus and related words, get influenced by other concepts from ontological prospects ... etc) .

2. Taking the previous point on consideration such pages tags will be indexed in G.A.S.E.T system, the comparing task will be as easy as comparing words in the standard keywords indexing method., ultimately Results will be superior to the one received by Google ™ search engine, because it will be based on more proper factors.

G.A.S.E.T Enhancement for Google ™ " Similar Pages Function " - 2

Conceptually / Contextually Similar to the User Targeted Web Page

Target Web Page

CSFS

ANALYZER

ANLP

CCD

Add weight

Consulting

KEB:

1. LPGS
2. Analyzer
3. DERWS

... etc

ANLP:

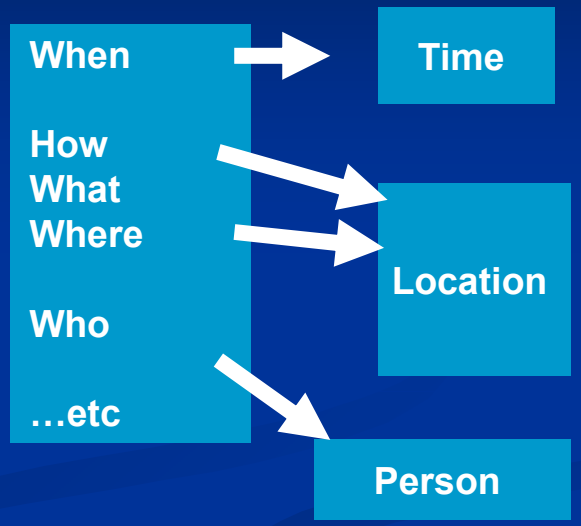
1. Concepts
2. Context
3. Contrast

... etc

Basic Structure of the G.A.S.E.T QA-System - 1

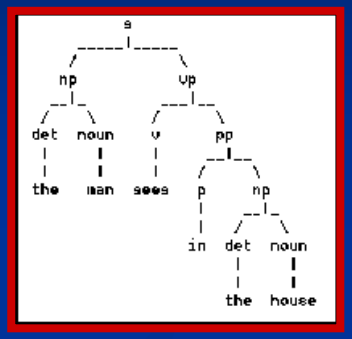
We where able to answer semi - complicated semantic queries using "unstructured" natural language sources. By implementing the following techniques:

- Analyze at the query related sense / concept in a multi dimensional way.
- Look for the logical respond to the linguistically and conceptual form of the query from our KEB .
- Parse the results according to the previous rules and to its own linguistic construction .



The GASET QA-System will be Fully Functional upon the Final Uploading of our project on Suitable servers with the needed QA / NLP defines Web Repositories and the powerful concept analyzing procedures which the system will eventually need to achieve its ultimate goals, such tasks will need Extensive and Detailed Explaining which we will provide if required.

Basic Structure of the G.A.S.E.T QA-System - 2



Query

Question

Question Formats Recognition

QA Conceptual Filter (Classifier)

Send

Type of Answer

Documents / passages

G.A.S.E.T ™ ANLP Defined Repositories

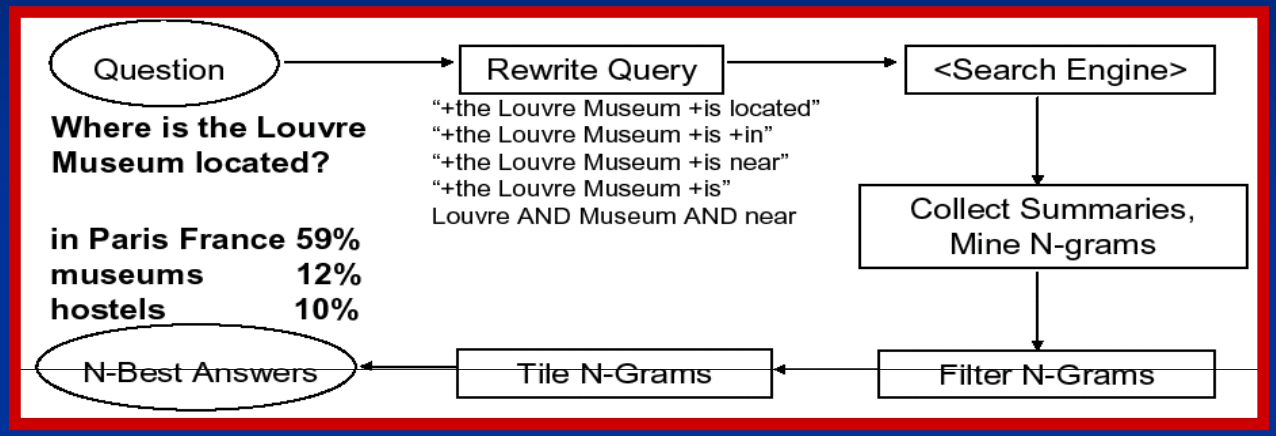
Proposed Answers Parsing and Verifying

Answer

Starting with the proper **semantically motivated** analysis will help eliminating tagging problems and greatly speed the results generating process

Basic Structure of the G.A.S.E.T QA-System - 3

Examples of other Tools



The old way and the G.A.S.E.T way:

New prospective and superior technology

Other search engines, which have some NLP capabilities, will try to match your question to their own Knowledge Base of Answers or "Rephrase" the question format in hope of finding sentences which will match the answer, such tactic miss the soul of QA, our specially tagged – analyzed QA web repository will dynamically find answers matching not only the context of the question but also its essence. It will also generate harmonized answers from compatible sources .

Thank you for your time

We appreciate your Queries and Comments . For more details and updated Technical Prospects, Financial Docs and Futuristic Potentials, please check our project dedicated website :

www.gaset-gbset.com

or contact, Mr. Wiam Gharbeyah at

weam@gaset-gbset.com