

Internet  
Data Management System

# Avalanche

"Non enim paranda nobis solum,  
sed fruenda sapientia est."

[" For wisdom is not only to be  
acquired, but to be utilized."

—Cicero, De Finib., i. I.]

© Inreco LAN, 2002

# Presentation plan

1. The problem to be solved
2. Description of needed software
3. The solution
4. **Avalanche** features and advantages
5. **Avalanche** detailed description
6. Instruments and technologies used
7. Installations

# Internet Surfers

There are two different Internet users groups having to fulfill the same task day by day.

## These groups are:

### This task is:

gathering and storing Web-information.

- Usual Internet users collecting information on the hobby they have (basketball news, cooking recipes, pets info, etc.)
- Analysts having the job of gathering and sorting Internet data (e.g. for Gartner Group, Bloomberg or IDC).

# Step-by-step Description

Let us list the steps to be done to solve the task mentioned above.

# Step 1 to solve the task

1. User needs to run some search or meta-search engine (e.g. Google, Yahoo, Copernic) and define the search query.

Let's keep in mind that different search engines have different syntactic rules for building the request and return very different results for the same request.

So, to make search more or less complete one needs to repeat it number of times for different search and meta-search engines with different syntactic rules to build the requests.

# Steps 2, 3 to solve the task

2. User needs to look through each screen of each output of each search engine thoroughly to filter only sites with the information that seems to be the one he looks for.
3. User needs to validate each of the filtered connections to understand whether they are alive or not.

# Steps 4, 5 to solve the task

4. User needs to enter each of the sites that have passed validation procedure and load its content to his local computer.
5. User needs to check few more links at each of the sites to load the content of the linked sites that is interesting for him.

# Steps 6, 7 to solve the task

6. After downloading all the needed data one has to make some steps offline. First of all he has to examine all the downloaded files thoroughly to place each of them to the corresponding subfolder in his usual file system folder organized to store files downloaded from Internet.
7. Now, to find any file by some keywords among the files being stored user could use only standard Windows search system of very limited abilities (no hyperlinks, no cookies, etc.).

# Conclusion

It was an absolutely honest description of the steps every user should do each day to get and use information he is in need of.

Using some helpful tools and hints (iHarvest software, Telnet software, MyYahoo module, schedulers, etc.) does not change the situation substantially.

Beyond any doubt the process of constantly repeating Web-search to find and store new information on given subject is too boring now. Something to be done with it. Let us describe how this process *should have been handled* if appropriate tools existed.

# How it should be. Step 1.

1. User creates special Smart Folder (if needed – with subfolders). Each Smart Folder has to gather from Internet and store only definite documents. User is provided with special means to describe not only usual syntactical, but also *semantic* requirements for the needed documents.

Thus, user describes *once* what new documents he wants to be *automatically* found in Internet on regular basis.

## How it should be. Steps 2 and 3.

2. User defines the schedule for automatic refreshing of Smart Folders (weekly, daily, several times a day).
3. User runs the system.

*That is all.*

After that, according to the schedule he had defined, user will find in Smart Folders new documents for the subject he is interested in. Everything else is up to system: syntactical Web search, checking links chains, downloading the documents, semantic analysis and classification.

# Special tool needed

Thus, nowadays the market is in need of software that would be aimed to do the following:

- Search for needed data through the Web on scheduled basis.
- Try found links and filter Internet content.
- Collect filtered data.
- Classify collected data.
- Store classified data providing the ways of flexible and comfortable access to stored data.

# Why there is no such software now?

Each of the existing software packages solves the problem partially (covering little part of the problem).

A software tool to solve the problem as a whole should be considerably complex. It should combine modules of substantially different functionality:

- Surfing Web and downloading Internet-content
- Classifying downloaded information
- Storing data with comfortable access to them

Complexity of some of these modules is usual programming complexity, and as for the task of classifying – it is not so easy mathematical task.



# We did it!

Inreco LAN company has developed a software system called **Avalanche**

**Avalanche** is an Internet Data Management System.

IDMS **Avalanche** contains number of new generation tools for:

- knowledge mining;
- knowledge storing;
- knowledge representing.

# Avalanche has number of competitive advantages

**Avalanche** - is an *inexpensive personal* tool that does not yield to the similar large corporate systems; moreover *Avalanche overcomes* a lot of them in power and in convenience.

- **Avalanche** contains *all the needed subsystems*: for direct Web-search, for customizable meta-search, for semantic classification, for convenient storage, for enhanced search through the stored documents, etc.
- **Avalanche** implements an idea of *Smart Folders* being automatically refreshed. Smart Folders are also adaptable to the scope of implementation.
- **Avalanche** makes it possible to make not only meta-search, but also to make *direct search according to preset IP ranges*. Thus user receives new documents instantly, before the appropriate sites are indexed by the standard search engines.
- **Avalanche** also can fulfilling *fast syntactical search* similarly to the way usual search machines do it. However, *Avalanche* shows to user only *new* entries as soon as it remembers the history of the previous searches.

# **Avalanche** is a single product with number of logically connected functions

- Syntactic and semantic definition of needed information.
- Means of scheduled data search in WWW.
- Semantic filtration and classification of incoming data.
- Means of creating user's personal encyclopedia.

**AVALANCHE**

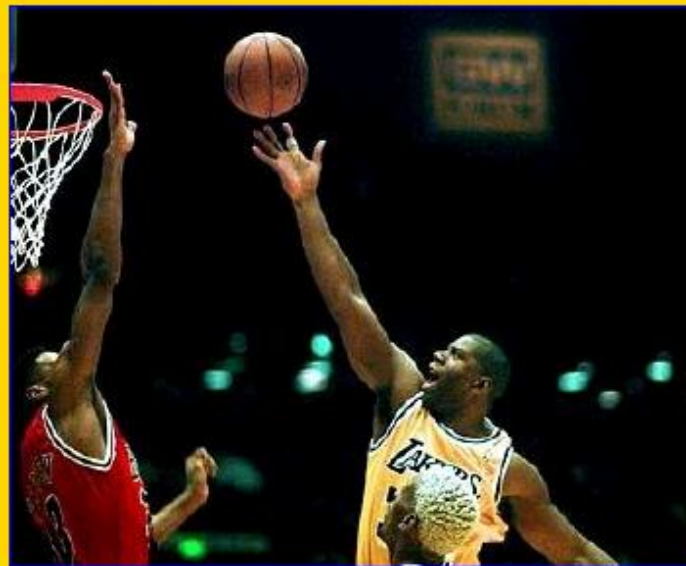
Catalogues Documents Actions Help

Catalogues	Meta-Catalogue	N°	Title	Received on	Sample	Unremovable	Relevance
[-] NBA Catalogue (2)		2	Magic Johnson	03.12.2000	<input type="checkbox"/>	<input type="checkbox"/>	
[-] NBA Catalogue (2)		1	NBA AT 50: MAGIC JOHNSON	03.12.2000	<input type="checkbox"/>	<input type="checkbox"/>	

NBA Catalogue (2)

- NBA Catalogue (2)
  - TEAMS
    - Los Angeles Lakers
    - San Antonio Spurs
  - PLAYERS (2)
    - Shaquille O'Neal
    - Magic Johnson 2
  - Trash

## Magic Johnson



Magic Johnson joined [Larry Bird](#) as the two marquis players of the 1980s. They were forever linked after Magic's Spartans defeated Bird's Indiana State team to win the NCAA title in 1979. Magic was later named tournament MVP. Magic was the

# Syntactic and semantic definition of needed information

**Avalanche** provides user with special means to create Smart Folders and to describe for each of them not only usual syntactical, but also semantic requirements to the documents to be gathered and stored in those folders.

Smart Folder is defined in terms of:

- presence or absence of certain words and phrases in new document;
- *computable proximity* of new document to number of sample documents.

Thus, changing key words and selecting maximally adequate sample documents, user can describe the search scope with *any level of detailing*.

# Example of syntactic and semantic definition

The screenshot displays the AVALANCHE software interface. The main window shows a tree view of catalogues under the 'Meta-Catalogue' tab. The 'NBA Catalogue' folder is selected, and its 'Folder Attributes' dialog box is open. The dialog box has two tabs: 'Main' and 'Advanced'. The 'Main' tab is active, showing sections for 'Key words/word-combinations' and 'Forbidden words/word-combinations'. In the 'Key words' section, the 'Additional' list contains 'NBA Catalogue'. The 'Logical expression' checkbox is checked, and the text field contains the expression: 'NBA OR Los Angeles Lakers OR Lakers OR San Antonio Spurs OR Shaquille O'Neal OR Magi'. The 'Advanced' tab is currently empty.

**AVALANCHE**  
Catalogues Documents Actions Help

**Catalogues** **Meta-Catalogue**

- NBA Catalogue
  - NBA Catalogue
    - TEAMS
      - Los Angeles Lakers
      - San Antonio Spurs
    - PLAYERS
      - Shaquille O'Neal
      - Magic Johnson
  - Trash

**Folder Attributes**

Main | Advanced

Key words/word-combinations.

Additional	Inherited
NBA Catalogue	

Forbidden words/word-combinations.

Additional	Inherited

Logical expression:  
NBA OR Los Angeles Lakers OR Lakers OR San Antonio Spurs OR Shaquille O'Neal OR Magi

OK Cancel

Total: 0, including new: 0, unread: 0, unremovable: 0, sample: 0

# Means of scheduled data search in World Wide Web

**Avalanche** includes Internet Spider that provides:

- scheduled *automatic search* of requested information in the Web;
- *automatic links following*;
- *automatic validation* of found links;
- *copying* of found information from Internet to the user's local computer.

# Example of scheduled data search

Search Engine Name	Search Engine Address	Use This Search Engine	Use Links Limit	Links Quantity Limit
Google	http://www.google.com	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	50
AltaVista	http://www.altavista.com	<input checked="" type="checkbox"/>		
Yahoo	http://www.yahoo.com	<input checked="" type="checkbox"/>		
GO	http://www.go.com	<input checked="" type="checkbox"/>		
Lycos	http://www.lycos.com	<input checked="" type="checkbox"/>		
WebCrawler	http://www.webcrawler.com	<input checked="" type="checkbox"/>		
HotBot	http://www.hotbot.com	<input checked="" type="checkbox"/>		
Excite	http://www.excite.com	<input checked="" type="checkbox"/>		
AOL	http://search.aol.com	<input checked="" type="checkbox"/>		
Direct Hit	http://www.directhit.com	<input checked="" type="checkbox"/>		
CNET	http://www.search.com	<input checked="" type="checkbox"/>		

### Search Engines Settings

Actions View

Search engine: Google

Use this search engine

Use links limit

Links quantity limit:

Search engine name:

---

Links Page links

Chosen links	Similar links
http://www.audi.de/index_de.html	http://sports.sleuth.com/team-non.cfm?
	http://www.sports.sleuth.com/team-non.
	http://cnnsi.com/basketball/nba/boxsc
	http://cnnsi.com/basketball/nba/boxsc
	http://www.usatoday.com/sports/scores
	http://www.usatoday.com/sports/scores
	http://www.thefantasysportsribune.com.
	http://www.thefantasysportsribune.com.
	http://nba.koti.com/pl/his_mvp.html

When selecting links, take into account

Type name  Type style  Type size

### SPIDER

Current Statistics

Volume of	Value	Progress
downloaded information, total, Kb	868,89	<div style="width: 100%;"></div>
documents prepared for rubrication, Kb	242,57	<div style="width: 100%;"></div>

Amount of

links found	Value	Progress
links found	149	<div style="width: 100%;"></div>
links processed during session	10	<div style="width: 100%;"></div>
links rejected during session	0	<div style="width: 100%;"></div>

Documents prepared for rubrication: **4**

---

Current Processes

Process	Process Information	Process Status
Search Engine Querying	Yahoo processed 1 pages	<div style="width: 8%;"></div> 8%
Document Processing	idle	
Document Downloading	idle	
Document Downloading	idle	
Document Downloading	http://membres.tripod.fr/nba_simulation/	downloading
Document Downloading	http://www.clarin.com/diario/2000-04-07/r-07102d.htm	downloading

---

Search Engines

Query: NBA OR Los Angeles Lakers OR Lakers OR San Antonio Spurs OR Shaquille O'Neal OR Magic Johnson

Search engine: Yahoo

**69%**

Process started... 0:01:53 03.12.2000

When selecting links, take into account

Type name  Type style  Type size

com/search?q=NBA%20OR%20Los%20Angeles%20Lakers%20OR%20Lakers%20OR%20San%20Antonio%20Spurs%20OR%20Shaquille%20Neal%20OR%20

[Advanced Search](#) [Preferences](#) [Search Tips](#)

OR Los Angeles Lakers OR L

ip: In most browsers you can just hit the return key instead of clicking on the search button.

c" (and any subsequent words) was ignored because we limit queries to 10 words.

es Lakers OR Lakers OR San Antonio Spurs OR Shaquille O'Neal OR Results 1 - 10 of about 164. Search took 2.39 seconds.

### Send to Spider

You are going to send the following information to Spider:

Query: NBA OR Los Angeles Lakers OR Lakers OR San Antonio Spurs OR Shaquille O'Neal OR Magic Johnson

Starting Addresses:

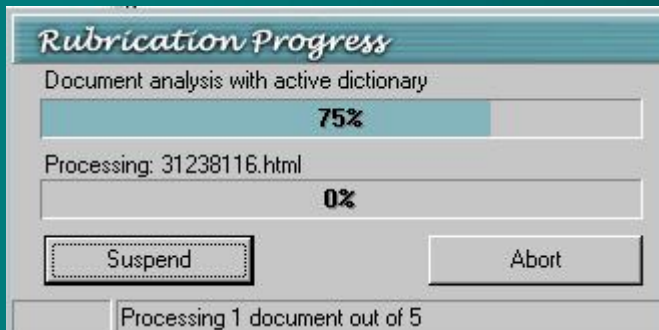
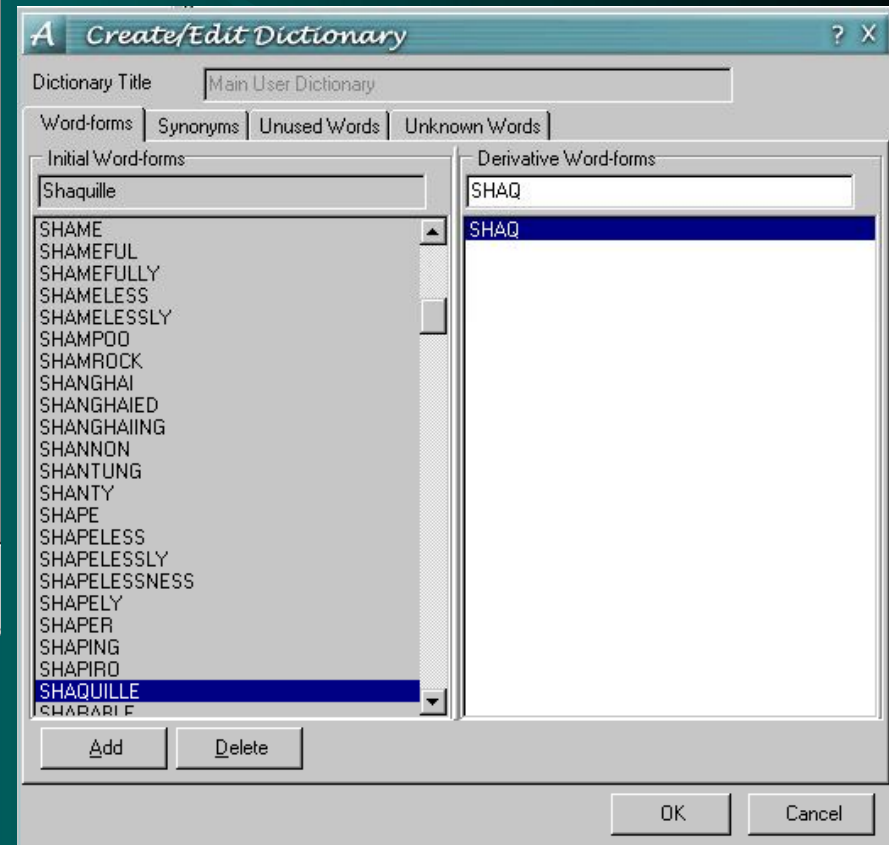
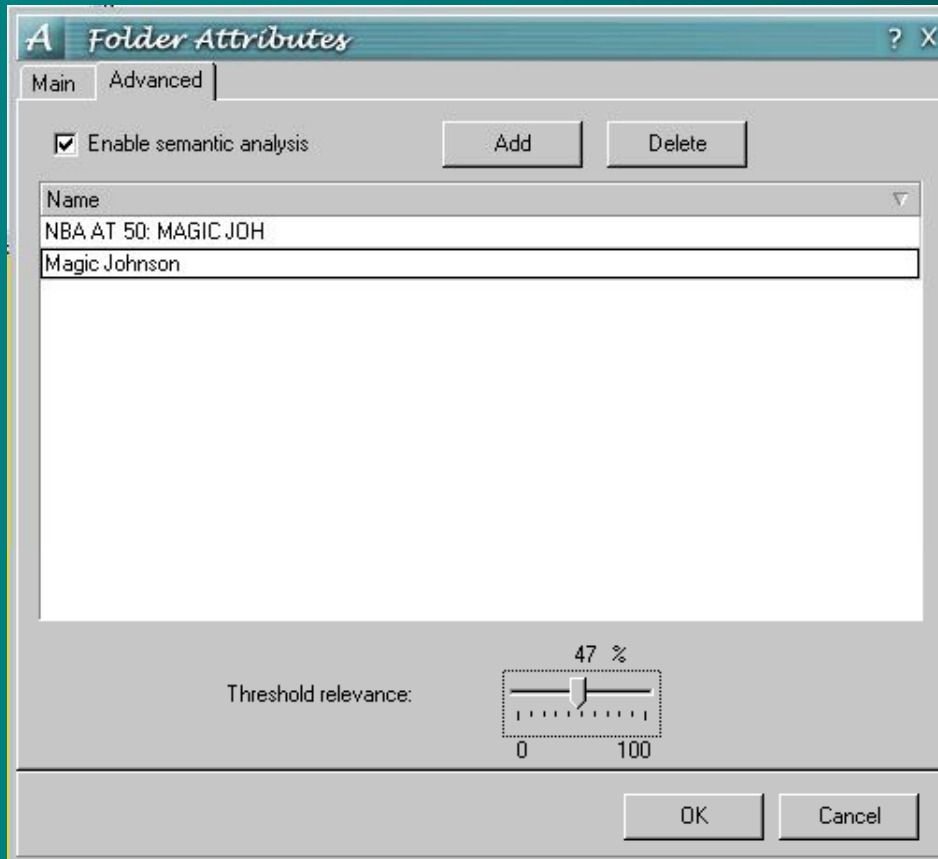
Spider will search documents in Internet using this information for catalogue "NBA Catalogue". Spider will also send this query to selected search engines.

# Semantic filtration and classification of incoming data

Our own algorithms of semantic classification have been implemented in **Avalanche**. They provide:

- *Automatic classification* of copied information in accordance to the Smart Folders structure.
- *Storage* of classified information. Information is stored on the local computer in efficient way.
- *Re-classification* of stored information. You can change your mind and reclassify information already got from Internet.

# Example of semantic filtration and classification



# Means of creating user's personal encyclopedia

**Avalanche** provides creation and management of *user's personal encyclopedia* built as a local Internet site for adequate description and convenient maintenance of stored information.

# Example of creating user's personal encyclopedia

Microsoft Internet Explorer

Address: C:\Program Files\Avalanche\Site\Index.html

# AVALANCHE

Home Meta-Folder Meta-Folder Meta-Folder Meta-Folder

- NBA Catalogue
  - TEAMS
  - PLAYERS
    - Shaquille O'Neal
    - Magic Johnson

Title	Received on	Parameters	Card
<a href="#">NBA AT 50: MAGIC JOHNSON</a>	03.12.2000	- -	<a href="#">Document Card</a>
<a href="#">Magic Johnson</a>	03.12.2000	- -	<a href="#">Document Card</a>

document: - new - sample - unremovable

Create HTML Encyclopedia

Cancel

Relevance: 0  
Expression: Magic Johnson

# **Avalanche** is flexible and scalable product

**Avalanche** could be fitted either for expert's analytical work or for common user's Internet surf.

# Instruments and technologies

**Avalanche** algorithms for data classification and texts proximity evaluation are developed on *strong mathematical basis*.

**Avalanche** is developed with *proven technology* that means following *standards* for all stages of project maintenance, programming and testing.

# Instruments and technologies

Different parts of **Avalanche** were designed and developed using most up-to-date and efficient tools and algorithms.

User interfaces were developed using Borland RAD tools. Core code is written using object-oriented approach which makes **Avalanche** highly configurable and flexible.

Class design was made using Rational Rose tools, which is considered to be the best OOP-design tool nowadays.

Database is designed and optimized to Normal Form III, that's why data are stored efficiently, without any redundancy. Data integrity is declared and applied on database level.

Dictionary and document searching is optimized by using latest hashing and caching algorithms combined with direct dictionary access.

# Installations

No special marketing activities has been held to sell **Avalanche**. However it has been already installed and is actively used in Harvard JFK School of Government and in number of large Moscow companies (in analytical departments of the companies).

Lightened version of Avalanche has been installed recently as a payable service at one of the biggest Russian ISPs MTU-Intel (by request of MTU-Intel).