Petrarch

Intelligent Web Search Agent

Version 1.0.1

DRAFT - Subject to Change - DRAFT

Reticular Systems, Inc.
4715 Viewridge Avenue, Suite #200
San Diego, CA 92123
(858) 279-9723
http://www.reticular.com

May 21, 2001

:

Reticular Systems, Inc.
4715 Viewridge Avenue, Suite #200
San Diego, CA 92123
(858) 279-9723
http://www.reticular.com

Introduction

Petrarch is an intelligent search agent for finding, analyzing and displaying web-based documents. Petrarch is unique in that it learns your information needs and adapts its behavior to find the best documents that are most suitable for you. The more you use Petrarch, the better documents it will find for you.

Petrarch Functions

Petrarch is designed to provide you with a number of services to help you find, analyze, and retrieve web-based documents. These services are described below.

- **Find** - Petrarch uses meta-search techniques to find documents that contain information of interest to you. This means that Petrarch is using multiple internet search engines to find candidate documents meeting your requirements.
- **Retrieve** - Petrarch uses information from the search engines to download documents of interest.
- **Analyze**- Petrarch uses advanced information retrieval techniques to analyze retrieved documents. Petrarch analyzes documents using the following steps:
    - ***Lexical Analysis*** - punctuation, capitalization and hyphenation are removed.
    - ***Stop Word Removal*** - common English words that do not affect the information content are removed.
    - ***Word Stemming*** - word endings such as -*ing* and -*ed* are removed to find the root of the words in the document.
    - ***Characterization*** - a document is characterized as existing in an n-dimensional information space. That space is determined by the unique words and frequency of occurrence of words in this information space. The characterization process generates a number describing the document in this n-dimensional space. When you rank or rate docu-

:

ments you are telling Petrarch where documents that interest you can be found in this space.

- **Organize** - Information is organized for subsequent display or storage by Petrarch based on how the document rates or ranks when compared with other documents retrieved and analyzed by Petrarch.
- **Store** - Documents can be stored in a local database on the Petrarch server. Search history and documents are stored across session. Future versions of Petrarch provide for local storage of retrieved documents.
- **Display** - Petrarch displays documents in two ways. You can view the raw document with HTML tags and Javascript removed to get a quick idea if this document is of interest. You can also view the actual retrieved document using your web browser.
- **Monitor** - Petrarch can continuously monitor a web site looking for specified information.
- **Notify** - When Petrarch finds information that it thinks will be of relevance to you it can notify you by E-mail.
- **Spider** - After Petrarch finds a page of relevance to you, it can spider all of the links on that page. This means that Petrarch can look at all of the pages referenced by the found page.
- **Ontology View and Edit** - Knowledge about a particular search domain is organized in a structure called an *ontology*. Ontologies are further defined in later sections of this document. Petrarch provides tools for constructing, viewing and editing ontologies. You can have multiple ontologies and ontologies need not be related.

Figure 1 shows a functional block diagram of Petrarch.

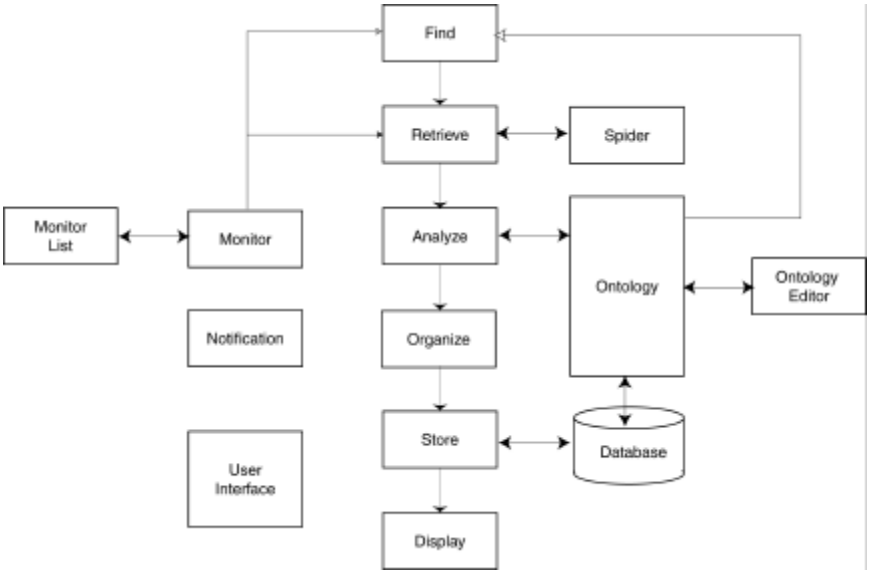

Figure 1. Petrarch Functional Block Diagram

Petrarch Operation

Petrarch operates a little differently that the internet-based search engines that you are probably familiar with. The Petrarch search agent needs to interact with you in the initial phases of your search activity so it can learn about the documents you desire to see. You will need to enter a set of keywords relevant to your search interest. Petrarch will then use existing web-based search engines to find a candidate set of documents that likely meet your needs. You will then be asked to rank or rate those documents and inform the agent how useful these documents are to you. Once Petrarch determines what you like, you can then ask it to go find more documents.

:

Ontologies

Petrarch uses something called an *ontology* to assist you in organizing your information and then finding information of relevance. An ontology is simply a way of structuring and viewing your knowledge about a particular concept. Thus, the formal definition of an ontology is that it is the *specification of a conceptualization*.

Ontologies in Petrarch are organized as simple tree structures. This is useful because nodes in the ontology can inherit information from their parent nodes. Figure 2 illustrates a simple ontology for the concept *intelligent agent*. Note that there is no single "correct" ontology for a concept. You define ontologies in a way that makes sense to you. Further, you determine how you want to organize your knowledge. However, you will learn as you use Petrarch how the structure of the ontology can aid in finding information.

Metasearch

Petrarch uses Metacrawler metasearch for finding candidate documents. This means that rather than using a single search engine, you are using many search engines to find information. The current version of Petrarch uses the search engines shown in Table 1.

Table 1. Metasearch Engines

AltaVista	DirectHit
Excite	FindWhat
Google	GoTo.com
Internet Keywords	Kanoodle
LookSmart	Lycos
MetaCatalog	Sprinks by About
WebCrawler	

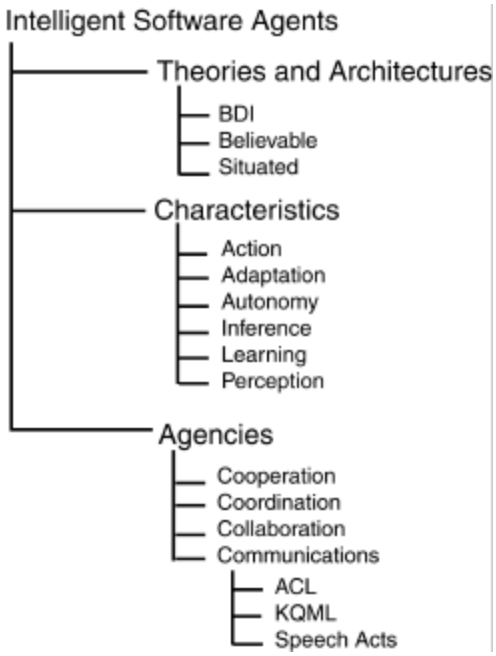

Figure 2. An Example Ontology

Some search engines do not necessarily return the same results every time you perform multiple searches using the same keywords. Also, sometimes one or more search engines may be off-line because of maintenance or network congestion. This means that Petrarch will not necessarily return the same results in every session where you look for documents using the same set of keywords in your ontology.

Using Petrarch

The following paragraphs explain how to use Petrarch. In this example, you will be building an ontology and looking for information about the planet *mars*. You must first log into the system by

:

entering your user ID and password. Once you have logged in successfully you will be presented with the dialog shown in Figure 3.

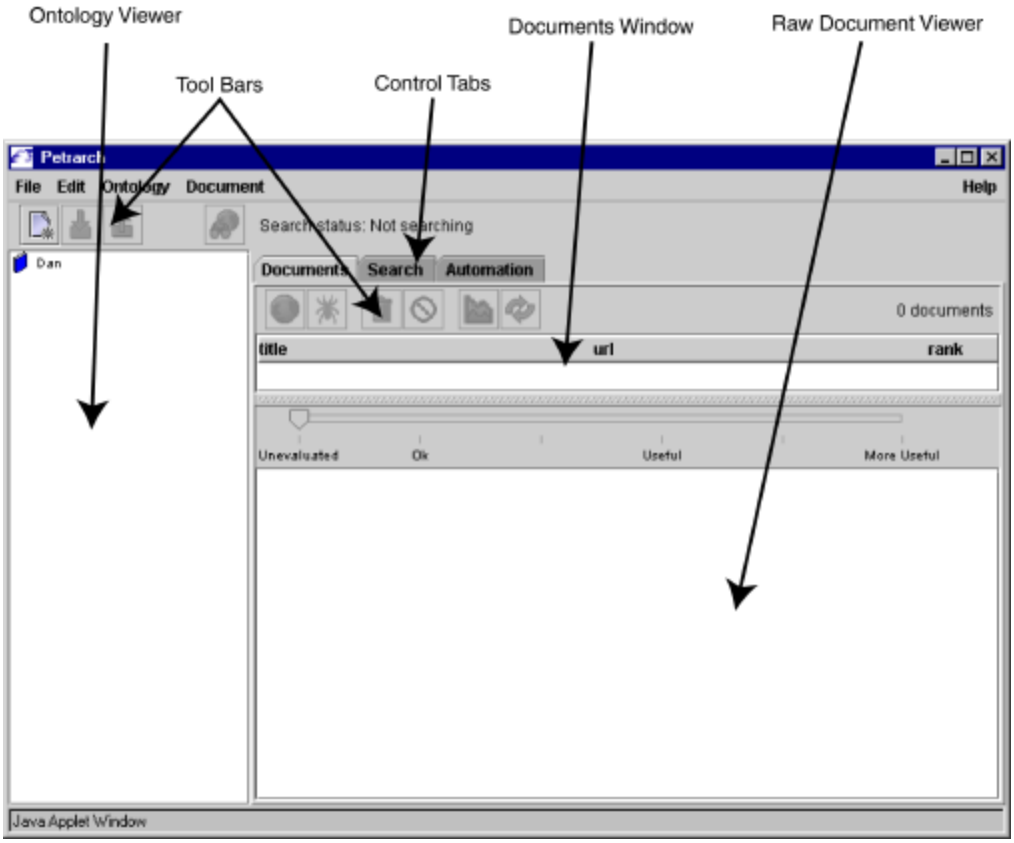

Figure 3. Petrarch Dialog

The **Ontology Viewer** area shows the ontology for the selected user. Note that the initial view defines an ontology with the username of the user. You can change this if you wish.

The **Documents Window** is a panel used to display information about the documents retrieved by Petrarch. You can resize this panel (or any of the other panels in the dialog) by moving your mouse over the dimpled bar above the evaluation slider. When your cursor turns into a vertical line with arrows on each end, click on the bar and drag to resize.

The **Raw Document Viewer** is used to display a retrieved document in its raw form. This means that HTML tags, Javascript, images and other extraneous information are removed. You can use this window to inspect a document and see if it might be relevant prior to ranking it.

The **Control Tabs** are used to control the operation of Petrarch. You can uses the tabs to select a view of the **Documents**, or to set up keywords for **Search**, or to set up **Automation** features (monitoring and notification).

Tool Bars are provided to perform various operation in Petrarch. The various tools are described in the following paragraphs. We are now ready to create a simple ontology and initiate a search.

1. If you don't have an ontology, select the **Ontology → New Ontology** menu item to create a new ontology.
2. Select the root of your ontology. (This is the icon that looks like a book). You may need to double-click this Icon to open it and see the ontologies stored in it.
3. Select **Ontology → Add Concept** menu item to create a new concept in your ontology.
4. A dialog box will appear. Name the concept *mars* and click **OK**.
5. Select the **Search** tab. You can modify the text in the **Search Keywords & Phrases** text area if you desire.
6. You now need to set the Minimum Ranking Threshold. For now, set the **Minimum Ranking Threshold** to 15 by dragging the slider.

:

7. Set the number of documents to be retrieved to 10. Do this by clicking on the **Retrieve up to *n* Documents** combo box and selecting 10.
8. Click on the **Documents** tab.
9. Select the **File → Search...** menu item to start Petrarch searching for documents. (You could also use the **Search** toolbar button that looks like a pair of binoculars). You will be presented with a dialog box titled **Friendly Message** informing you that you do not have enough documents to perform document ranking. This is expected because this is your first search.
10. Press the **OK** button in the dialog.

Note that the search status is displayed next to the **Start Search** toolbar button. The search process may take longer than traditional search engines, because Petrarch is processing and analyzing documents in real-time, not just returning the links that the search engine has previously indexed.

11. Wait until the search status area displays the *Not Searching* in its status field.
12. When the search has completed, the document list will be refreshed in the **Documents Window** portion of the panel and will show the documents that have been retrieved.

If you click on a document in the Documents Window, its contents will be displayed in the Raw Document Viewer panel. When you click on document to view it, its color changes from green to black to indicate that it has been examined. Note that Petrarch will likely not retrieve exactly the same documents and the documents retrieved will not necessarily be displayed in the same order as shown in the example.

Figure 4 shows Petrarch with the results from a search using *mars* as the keyword. Note that “NASA store...” is a site for mars memorabilia and is not really what we are interested in. However, the

mars-watch.com site appears to have relevant information. You can verify this by selecting a document in the document list and selecting the **Document → Open in browser** menu item.

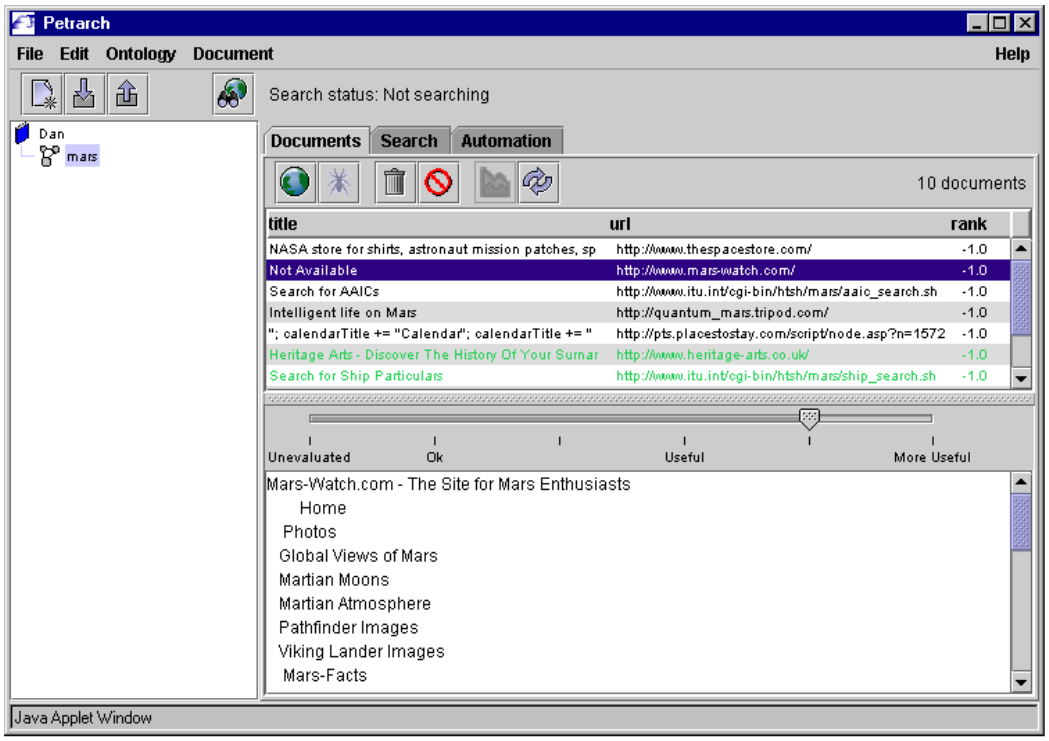

Figure 4. Petrarch Retrieving *mars* Documents

13. Now you need to evaluate the documents. When you see a document that is useful then set the slider to the appropriate position, e.g., halfway between **useful** and **more useful**. Review each document and evaluate by setting the slider. Documents that are not useful can be left at a setting of **unevaluated**. You can also delete (trash can tool) or ban (slash-circle tool) these documents as well.

:

14. Switch to the **Search** tab again and set the following properties:
    Minimum Threshold Ranking = 15.
    Number of documents to be retrieved = 20.
15. Repeat Steps 7-10 above.

Note that while the search is executing, you can select **Document ➜ Refresh Documents** as soon as the search status indicates that more than zero documents have been analyzed.

Note that all of the documents in the list are now ranked, including the documents that have been previously retrieved. Continue using Petrarch to retrieve, analyze and classify documents. You can also create new ontologies or add new concepts to your ontology.

You have now been introduced to the basic operation of Petrarch. Now you can perform basic document searches. Advanced features are also available including the ability to automate monitoring of web sites and spidering of documents. Please see the Petrarch documentation for details on these features.

Toolbars

The toolbar provides buttons for performing the following actions:

- Create a new ontology
- Add a new concept as a child to the selected concept
- Remove the selected concept
- Start searching for documents/Stop searching for documents
- The Documents tab toolbar contains the following buttons, in order:
- Open the selected document in a browser
- Spider the selected document (only 1)
- Remove the selected document
- Ban the selected document (from future retrieval for this concept)
- Re-rank all documents

- Refresh the document list