



<http://www.viceversatech.com>

The Smart Content technology

Growing amount of information dispersed across different sources is an increasing problem of state-of-the-art information management. To be solved effectively, it demands new approaches and tools, strongly focused on the content semantics and supported by automation and intelligence.

SCAN (Smart Content Aggregation and Navigation) technology combines semantic integration, search and natural language processing for intelligence of document management in the age of information overload. In order to provide new, improved experience for knowledge workers, the technology addresses a broad range of major issues and challenges of state-of-the-art information management:

- The information is dispersed across different heterogeneous sources and locked into numerous application-specific formats. It sets a barrier for a high-level overview of existing knowledge base and findability of semantically related information pieces.
- The structure of information is mostly dictated by a physical storage technology (an example - files and folders in the filesystem), while more high-level, semantical structures are needed.
- There is no uniform way to describe and annotate the content resources, supported consistently across a whole system. Even if a document format supports metadata, it does not contribute into overall information management environment and is mostly useless outside an application that uses this format.
- Despite of the computers are smartest of all artificial things now, there is no pro-active help in the content classification and information modeling from their side. The computers role in information management still to be the passive storage and processing systems.

SCAN technology is highly adaptive and totally unobtrusive. It does not require a special content management infrastructure or changing your business processes in order to using it. You will not have to change the way you work with the information day-to-day – you still can store your documents where you are used to, send and receive emails, bookmark your favourite websites and so on. SCAN does not touch your content either, but adds a layer upon it to turn your content to the *smart content* - that is, the content that *knows something about itself*.

An integrated approach

SCAN approach is based on an idea that information overload is a complex of problems and no single magic bullet exists to solve them effectively. Thus, there is a synthesis of few different techniques aimed at specific aspects of the problem.

Content aggregation

SCAN erases the boundaries put on information by different storage systems. It links the information items from multiple sources and of different formats into a seamless digital library, where they can be categorized, annotated navigated and searched by a uniform way. This provides a homogeneous searchable and explorable semantic information space where files, web-pages, emails, other content items are equal documents organized by their natural semantic properties.

The component architecture makes the technology agnostic of specific types of sources (local or network file systems, web-sites etc.) and of the document formats (MS Office, PDF, HTML ...). A number of those types of sources and formats can be supported via integration of the components for a specific business application.

Tagging

Tagging is the easiest and intuitive way of information modeling and organization of the documents collection similarly to the popular services like del.icio.us or Flickr. Tags are keywords or labels freely attached to the items to identify them for quick navigation and finding. All tags together form a taxonomy representing the semantics of the documents collection. The taxonomy can be viewed as a “tags cloud” for navigating through the documents.

Text analysis and concept extraction

SCAN brings the power of automated text mining and natural language processing to discover document semantics by extracting the valuable terms and their patterns from the document content. It makes possible to identify what the document is about and how it relates to others.

Text analysis greatly simplifies the process of tagging. It helps a user to pick the most relevant terms identifying a document and assign them as the document tags. It makes the manual document tagging as simple as selecting the tags from the

suggested candidates. Also, a user can entrust the process of tagging entirely to the system, so that the documents would be tagged automatically with the relevant terms.

Text analysis is also underlying the advanced semantic search functionality like finding the documents by similarity (pattern search) and associative guided search based on system suggestions.

Metadata and facet navigation

SCAN provides a rich set of metadata properties associated with the documents, including document title, description/annotation, author, creation date and others. The properties are set automatically on a document adding and can be quickly edited later.

Metadata properties can be used in the structured search to find the documents matching specified criteria. In addition, some properties (e.g., author or creation date) may be used as navigation facets to browse the documents collection.

Search

The documents content is indexed for search - either unstructured full-text search or more complex, structured queries both on text and metadata properties. It is possible to save the search queries for repeatable use, thus creating the dynamically populated sets of the documents grouped by specified criteria.

Advanced semantic search techniques are driven by text analysis functionality. After any search request is performed, the results are analysed to build a “see also” terms list for associative search. It provides step-by-step exploration of an area of interest following the system recommendations. Pattern search explores semantic compatibilities between different documents and allows to find the documents conceptually similar to a subject of a given one.

Implementations

We are confident that every business information environment starts from a personal desktop of an individual worker. So, the first implementation of the SCAN technology was a desktop system for personal information management. SCAN Desktop is a full-featured stable platform-independent product available for free under an open source license.

We are thinking also about business-oriented applications of the SCAN technology and our current R&D efforts are focused on design and development of the server version for small and medium enterprise networks.

For more information

You can read more on SCAN technology at our company web-site: <http://www.viceversatech.com> and on the SCAN Desktop community web-site: <http://scan.sf.net>.

* * *

ViceVersa Technologies is a software development, research and consultancy firm, aimed at delivering an integrated semantic search, text mining and metadata solutions to empower personal and corporate information management by innovative semantic technologies.

ViceVersa Technologies is always interested into partnership and cooperation with commercial companies and academic institutions for collaborative research and development projects.

ViceVersa Technologies
Komvuzovskaya st. 21-G, 322
Ekaterinburg
620066 Russian Federation

Phone: +7 343 383-2338
E-mail: info@viceversatech.com