# What is Contextual Search?

By Bjørn Olstad, CTO, and Silvija Seres, Senior Director, Corporate Development
Fast Search & Transfer, Inc.™ (FAST™)

Contextual search is the convergence of three dominant approaches within the major areas of modern information retrieval:

1. Intelligent text and data mining, enabling information discovery through automated concept modeling and exploration of patterns;

2. Flexible content structuring technologies, enabling XML schema independence and cross-connection of content from structured and unstructured sources; and

3. High-performing search technology, enabling scalable and fast content gathering, processing and retrieval in face of exploding information volume and complexity.

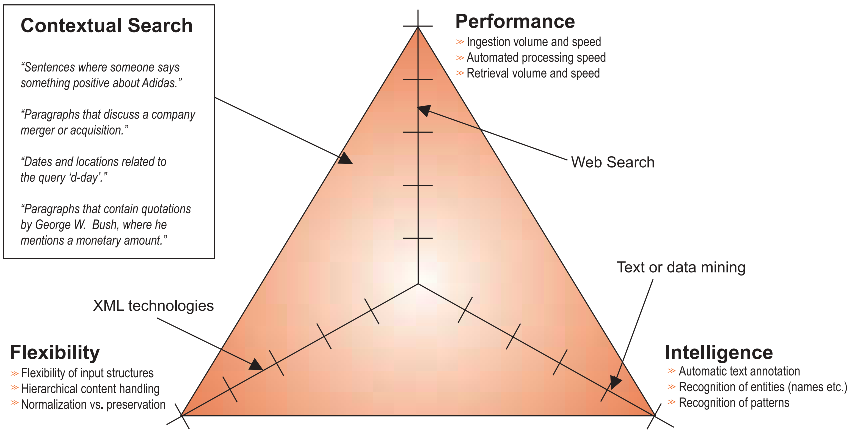
Text and data mining are the automated discovery of new information by extraction of patterns and relationships between entities in text or structured data sources, respectively. They can, for example, be used to discover new patterns in consumer behavior, or help research in biosciences by exploring implicit links between concepts in scientific texts. The discovery is even extended to rich media.

With the advancement of XML technologies, a de facto standard document structuring framework allows authors to define their own sets of tags and document structures, also known as schemas. Retrieval systems dealing with a large number of sources need the same flexibility—they must be "schema independent." Contextual search engines provide this independence by replacing predefined index layouts with a nested structure that has scopes and tags. With this, new types of precise queries can be asked that combine structure and content, imposing contextual constraints on the content. For example, the users can formulate queries in XPath or an optimal subset of XQuery.

Scalability and consistency of the system are the very foundation for heavy data crunching behind the scenes, seamlessly filtering and improving structured, unstructured and rich media content, queries and results, with no performance penalty for the users. This power depends on simultaneous scalability in several dimensions: data volume, query traffic, data and query complexity, fault tolerance, real-time capabilities, etc.

## New Advances

The combination gives unprecedented freedom to both front-end providers and end users. Traditional systems allow only "index time content mining" (ITCM), where data models are defined before indexing time; after indexing, the original content is typically discarded, so important relations built in the context are irretrievably lost, and the discovery is based on models that are on average 80% correct. Contextual search also provides "query time content mining" (QTCM), allowing end users to be content miners, on structured, unstructured or rich media content—an area previously reserved for deep experts. Contextual search combines ITCM, used for entity extraction and other content enrichment, with QTCM, used to find new relationships at query time based on the structural information preserved in the indexed content. The embedded contextual metatags are essential in mining for facts, since global metatags lose valuable implicit information.


Contextual Search
"Sentences where someone says something positive about Adidas."
"Paragraphs that discuss a company merger or acquisition."
"Dates and locations related to the query 'd-day'."
"Paragraphs that contain quotations by George W. Bush, where he mentions a monetary amount."
Performance
≫ Ingestion volume and speed
≫ Automated processing speed
≫ Retrieval volume and speed
Web Search
Text or data mining
XML technologies
Flexibility
≫ Flexibility of input structures
≫ Hierarchical content handling
≫ Normalization vs. preservation
Intelligence
≫ Automatic text annotation
≫ Recognition of entities (names etc.)
≫ Recognition of patterns

Contextual search can also be viewed as a bridging technology, merging search engines and databases. These technologies were originally developed with different data in mind (unstructured vs. structured), but are rapidly approaching each other in terms of data and feature capabilities. However, some inherent differences continue: search engines are built to find enough good results fast, while databases are built to find all results but with less consistent speed. The long tail of database retrieval distribution shows the slowest 20% of queries with multi-second responses times. With the emerging use of unstructured or rich data by databases, the long tail is bound to get even longer. Contextual search provides a hybrid solution that provides the best of both worlds—it consistently cuts the retrieval time to sub-second, while keeping the completeness of result set, and providing a combination of rich relevancy ranking models, multi-field sorting, advanced linguistic models and joins. It is also used as a complement to databases, for database offloading, with the workload split by allowing database technology to power transactional processing and contextual search to power retrieval and mining.

## Scope Search

A new retrieval capability arising from contextual search is "scope search." Here, searches can be restricted to structural parts, or scopes, of documents, such as sentences, paragraphs or any entity of interest, rather than to whole documents as is usual in search systems. Documents are no longer the atomic retrieval units, and unstructured content becomes structured. The scope units may be structural and semantic regions in text documents, numeric data tuples in databases, or speech parts in audio files. These elements are marked in context, where grammars act as scope producers, and scopes are annotated as XML local metadata. This is fundamental, since it preserves the context of the original information for later searching, and makes it possible to act on all of the information.

For example, with scope search one can match documents where the same sentence contains a company name and the word "scandal," and also aggregate all personal names that occur in sentences that contain the job title “CFO.” Other examples of possible contextual searches are:

- Paragraphs that contain quotations by George W. Bush, and mention a monetary amount;
- Sentences where someone says something positive about Adidas;
- Dates and locations related to the query “d-day.”

In effect, with scope search, all relevant text and other data is automatically categorized and set in context. Before, names of people, names of places, dates, prices, scientific concepts were blended together and effectively lost in the index; now, they are related facts waiting to be exploited. The richness of its surrounding context allows valuable information to be separated from noise, thus allowing factual discovery at unprecedented speed and ease of use.

| *Persons that appear in* **documents** *that contain the word “soccer”* | *Persons that appear in* **paragraphs** *that contain the word “soccer”* |
|---|---|
| **person@base** | **person@base** |
| Jack Nicklaus (~10.0%) | Diego Maradona (~4.0%) |
| Fred Davis (~10.0%) | David Beckham (~4.0%) |
| Billie Jean King (~8.0%) | Alan Shearer (~3.0%) |
| Richard Nixon (~7.0%) | Michelle Akers (~3.0%) |
| John Wayne (~7.0%) | Mai Hamm (~3.0%) |
| Margaret Smith (~7.0%) | Eric Wynalda (~3.0%) |
| Joe Frazier (~7.0%) | Freddy Adu (~3.0%) |
| Irina Rodnina (~7.0%) | Michel Platini (~2.0%) |
| Mao Zedong (~6.0%) | Stanley Matthews (~2.0%) |
| Gordie Howe (~6.0%) | Olivier Neuville (~2.0%) |
| Richard M. Nixon (~6.0%) | Bobby Moore (~2.0%) |

## More Examples

Document-level aggregation may yield too imprecise results, while restricting searches to sentences or paragraphs, or to some other constrained scope, provides a more factual and relevant relation between the query and the enriched content in the index. For example, in the sentence—*”In July, 2004, Leonid Kuchma was elected as Ukraine’s second president in free and fair elections”*—contextual search recognizes and stores the marked contextual units, respectively, as a date, name and location, in addition to the sentence itself. This semantic analysis and entity extraction allow queries such as *“When has Ukraine had elections?”* even though this fact has not been explicitly modeled. This is important—in the past, detailed semantic analysis systems have been only moderately successful exactly because such a priori modeling of all potentially interesting facts is hard. The contextual search approach reduces the need for a priori relationship modeling and instead identifies relationships on-the-fly at query time between user and content concepts.

Scope search also remembers how these units were related, and which facts these relations formed. In the sentence “*‘**Dynegy** has to act fast,’ said* **Roger Hamilton**, *a money manager with* **John Hancock Advisers Inc.**, *which sold its* **Enron** *shares in recent weeks.”* there are several facts stated: a person talking about a company, a person working for a company, and a company doing a financial transaction related to another company—this is valuable information that can be stored and re-discovered at query time, without any explicit modeling beforehand. In addition to data and text mining methods, the underlying tools combine grammatical processing, statistics and dictionaries to achieve this relational tagging. Natural language queries are also handled elegantly with this combination, where linguistic and statistical modules translate questions such as the one above, asking “*When was*” or “*Who was*,” and create appropriate queries, in more than 70 languages. Finally, scope search enables normalization of recognized entities, and provides miners with automated suggestions to broaden or refine their queries. Together, these tools allow end-users to discover, summarize and disambiguate their information needs in a simple and effective way.

As an example of improvement in precision, we have run two test queries against the online encyclopedia Wikipedia. See the example above. Results on the left are from query: *Persons that appear in* **documents** *that contain the word ‘soccer’* and on the right from query: *Persons that appear in* **paragraphs** *that contain the word ‘soccer’*. The improvement in result quality is striking, yet the first list of results corresponds to a Web search result that we accept as standard!

## Power of Contextual Search

Contextual search allows the information provider to preserve all the original information and spend less time annotating and classifying. To end users, it gives ease of use through better content and features, and creative freedom to ask questions the providers may not have planned for. It unlocks the full potential of information, as the value of information is not based on the ability to store it, but by the ability to *use* the information.

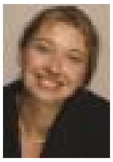
There are several essentially new aspects that contrast contextual search to other schema-flexible technologies. First, the retrieval performance is truly “industrial strength.” Second, data is weighted and filtered, rather than treated with an algebraic approach where all data has equal value, improving discovery quality. Third, retrieval can be done related to a large number of recognized entities, such as persons, companies or places, replacing result lists with answers. Fourth, new discovery tools are provided, including contextual navigation. Finally, contextual information vastly improves precision while preserving recall. Contextual search is not just about flexible query or content representation; it is about optimal content exploitation, turning information into value.

In summary, contextual search enables deep semantic analysis and refinement across structured, unstructured data and rich media, and dynamic interpretation of contextual meaning of the content. The overall result of the contextual combination is a vastly improved discovery, schema exploration and disambiguation capabilities. Bring on the next generation of searches! ▮

FAST SBP consultants are available to help your organization explore more effective uses of search-based solutions. FAST SBP provides these services in an objective fashion, looking for the best solutions for you. Please contact FAST SBP at sbp@fastsearch.com for more information, or visit us online at www.fastsearch.com.

Bjørn Olstad

Bjørn Olstad, Ph.D., serves as the chief technology officer at Fast Search & Transfer (FAST). Before joining the company, Dr. Olstad held key positions within General Electric Medical Systems, including director of research and development for cardiac ultrasound. He has served as a professor in computer science at the Norwegian University of Science and Technology (NTNU), where he was awarded the youngest professorship ever.


Silvija Seres

Silvija Seres has extensive scientific research background in algorithm design and optimization, with a Ph.D. and Prize Fellowship from Oxford University and ten international scientific publications and awards. Seres has held several long-term visiting research posts, including DEC SRC in the US, the National Science Academy in China, and a professorship from the first private female university in Saudi Arabia. Seres now works in Norway as a senior director of corporate development with Fast Search & Transfer.