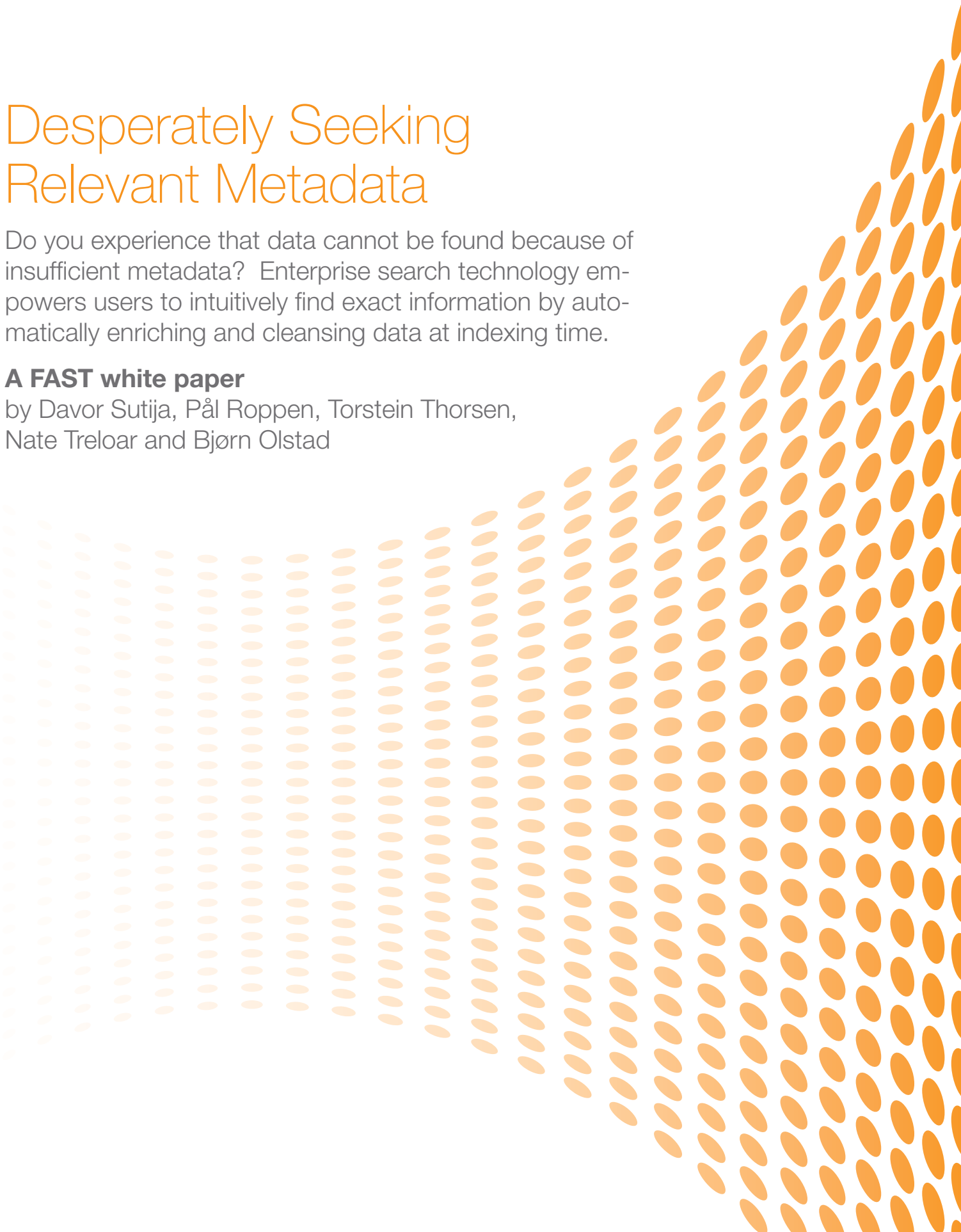
fast

Desperately Seeking Relevant Metadata

Do you experience that data cannot be found because of insufficient metadata? Enterprise search technology empowers users to intuitively find exact information by automatically enriching and cleansing data at indexing time.

A FAST white paper

by Davor Sutija, Pål Roppen, Torstein Thorsen, Nate Treloar and Bjørn Olstad



Introduction

The DBAs frustration often starts in identifying where to look for the information one needs. The problem becomes more complex when multiple repositories contain key pieces, but not the entire puzzle. How much time could be saved if one had a tool that identified, connected, and offloaded relevant information into a searchable repository with authoritative and complete metadata?

The solution that provides an answer to this question needs to be agnostic with regard to the type of data, and must access best-of-breed systems – both enterprise knowledge management and data warehouses. This paper will illuminate how knowledge management systems have evolved, how an enterprise search platform approach can broaden access to all company data by offloading databases, and thereby bring intelligence directly to business users.

Document Management is All About the Metadata

Missing, or incorrect, metadata is a significant problem. As an anecdotal example serves to illustrate the point, a census performed on our company's PowerPoint presentations a year ago showed that nearly a quarter were authored by the CEO himself. Doubtless because, as the founder, he created early presentations that have since been edited, copied, and modified many times over. As the saying goes: 'this is my grandfather's axe: my father changed the blade, and I changed the handle.' Lacking automatic systems that update metadata, the 'history' of the document can incorrectly categorize the current version.

It was precisely the problem with regard to metadata for unstructured information that led to the creation of knowledge management systems. In early versions, the metadata related to each 'document' was stored separately in a structured DB repository, and when information was searched for, a complex two-step process often ensued. First, a text search of the stored unstructured document matched the occurrence of keywords, while the same process had to be repeated on the metadata stored in the database. After two sets of documents were 'retrieved', the results were matched and the intersecting set was shared with the user. Needless to say, missing or incomplete metadata prevented relevant results from being returned to the user.

Data Warehousing Also Demands Complete and Cleansed Data

As for structured data, there was a time when the ideal was to create a single version of the 'truth' by creating appropriate data warehouses. While many strides have been made to make operational datamarts less important, the goals of bringing direct access to business users have not been fully met. And no wonder, with the size of data warehouses increasing by 5x every two years, and with the types of information required to be stored broadening to encompass increasing amounts of text within a variety of formats. Packaged apps, xml, and even html, now are to be found along with RDBMS in many warehouse environments. Combining their data is a growing headache.

In recent years, ETL tools have started supporting real-time data movement as well as incorporating certain data cleansing and profiling capabilities. Verification and data identification can be handled, and inconsistencies or redundancies detected, but their repair either relies on manual input or predefined dictionaries of names and addresses.

Enterprise-scale search provides an improved solution by converting data into relevant information at indexing time when data is first extracted and loaded, regardless of its degree of structure and independent of the underlying data model. Using linguistic tools, search technology can detect spelling errors, variations in orthography of names and geographical locations, use lemmatization (verb tense: run, running, ran) to identify word-forms, and most importantly, use relevancy models to score approximate matches. These may also assist data-cleansing of metadata, as authoritative sources with additional metadata can be then used automatically to supplement and cleanse incomplete or unreliable entries. The process highlights disambiguation of data using linguistic and logical tools, highlights and automatically corrects missing information, and integrates seamlessly with standard ETL tools.

Enterprise Search Enables Efficient Information Discovery

This approach to information access and retrieval is fundamentally different than conventional databases that rely on relational data models and join tables. While

ensuring referential integrity and transaction efficiency, an unintended side-effect is that relational databases are not optimized for "ad hoc" information requests wherein the end user can query without knowing the data model a priori. If the generated query requires complex joins or 'full table scans', or even if it is a query on a non-indexed field, a significant performance penalty is to be expected. A common complaint is that 'poor code' is the cause of performance problems (cf. DM Review May 2005); in our view, it is the need to provide 'good' code to ensure sub-second performance that is the challenge. Replacing an RDBMS with a search-based system allows the user greater flexibility and requires fewer knowledgeable inputs to produce a sub-second relevant result.

By using structured data, enterprises provide 100% sub-second response with distinct performance improvements – e.g. relevancy, navigators and presentation of comparisons between similar or synonymous items. A single structured search query can return 50 classes of variables with each instance ranked by occurrence, frequency, or range. A single query, with a typical response time of 100 ms delivers the equivalent of several thousand select and sort operations from a traditional database. In addition, the choice of navigators is easy to change - enabling enterprises to adapt to fast changing internal and customer requests. By adding more functionality and features, enterprises are enabled to improve the exposure and promotion of their products and services – visitors will simply experience richer and broader query results and be exposed to options they didn`t know existed. The advantage of a search-based system is that you do not need to set thresholds before being given an overview of the data distribution using navigators. The use of navigators is significantly more useful than choosing branches of a hierarchical tree structure, as each choice within a particular navigator is published with a corresponding frequency:

i.e. Job Openings = 1-4 (375), 5-10 (23), 11+ (3)).

The navigators can either be drop-lists of entities (companies, locations) or ranges (dates, frequencies, Corporate Sales Figures, etc). The ranges can be pre-assigned to be discrete units, or can be a set maximum number, with the distribution calculated at time of query.

Sales: <2,550,000 (24%, 36), 2,550,000-4,838,000 (25%, 37), etc.

When a given range is chosen, the navigator can be automatically recalculated to provide new buckets within the chosen range that automatically future refine the search. Thus, instead of making 5 predetermined choices, and guessing what lower limit of each variable to exclude, the user can simply type in a query, 'Java Programmer California', and obtain relevant guidance on what variables are useful differentiators, as well as relevant ranges.

Enterprise search technology is fundamentally designed to enable a user to request information *without* having to know a priori where the data lives, how it's structured, or even what attributes may exist to qualify the question.

Mini case study

E-commerce

A leading US automotive online marketplace with 11 million unique visitors per month and 3.2 million vehicles in inventory needed to redesign their website to accomodate increased customer volume and enriched functionality. They were looking for search technology providing real-time search performance and sustained query rates of 400 QPS. Their goal was to allow users to specify natural language queries such as 'Blue Car Miami <20,000' and have the system automatically find appropriate vehicles by matching query terms to classes of metadata and providing suitable navigators.

The choice of enterprise search was straightforward, in this case, as the need for extreme scalability and search speed was also compelling: RDBMS technology achieved only half the query load on the same hardware configuration. Implementation and full DB-offloading was completed in 12 weeks.

The benefits included:

- 100 % sub-second response
- Updating speeds of 500 documents per second
- Comparison shopping allowed similar offerings to be matched and alternative offerings presented (no Lamborghinis today, how about a Ferrari?)
- GEO search, offering vehicles within driving distance of the purchaser

Mini case study
Anti-Money Laundering

One of the leading government agencies in Europe actively detecting financial transaction patterns which indicate money-laundering and other illegal activity (false asylum seekers, organized crime) installed a search-based foreign exchange transaction monitoring system. All transactions over $4000 by foreigners in, international transfers, and credit-card transactions by citizens abroad are recorded. All of this data is stored using an enterprise search platform and can be searched securely by a variety of authorities with varying degrees of access. The National Statistical Bureau can extract aggregate reports on foreign exchange movements, while police authorities, and individual investigators, can use the system to investigate specific cases down to the level of individual transactions.

The speed, cost-effectiveness and navigation provided by Search supplanted an RDBMS-based design that was projected to cost more than twice as much. The system was operational ahead of schedule, and is expected to contain up to 10TB of information over the lifecycle of the program. A total of 50 navigators, (i.e. classes of metadata, such as customer name, transaction type, nationality, counterparty, date of birth, frequency of transactions) can all be displayed simultaneously, and the disparate reporting formats of each financial entity is seemlessly integrated into a common semantic and analysis framework.

It can be thought of as an additional abstraction layer that masks the complexity of the backend database, while providing high performance and ad hoc retrieval to the front-end application.

The user interface epitomizing ad hoc retrieval is a single search box wherein the user can enter a query and be provided options to refine their results. Rather than forcing the user to know, or be shown through a form, the entities, attributes, and relationships, they are instead provided an entry point into the data and offered options based on the actual data itself to narrow their results. This approach has the advantage of helping the user to discover novel patterns, trends, relationships, and anomalies in the data that they might not otherwise find.

Rather than constructing a data model that explicitly preserves the entity relationships in distinct tables, a search index "denormalizes" or selectively flattens the model across all relevant entities and their attributes. The entities and attributes exposed are driven by the application's search requirements, not by transaction performance or referential integrity requirements.

From Predictive Analytics to Contextual Insight

While there has been a lot of interest in predictive analytics that extrapolates data into the future, this has been previously limited to numerical data series. This can now be expanded by use of entity extraction that measures and scores the frequency of specific words and weighs their sentiment. In addition, enterprise search is now able to access the contextual nature of information, to break up documents into their logical structure, and identify whether in what sentence or report the match occurred, and determine from the surrounding context, its degree of relevance. As an example a query can be made to extract all *sentences* mentioning a product and living inside a paragraph that has a reference to a location, such as a store or geographic area. Using contextual insight, in combination with analytic tools, increases the precision of forecasts and KPIs.

Information discovery using Search is a proven technology

This technology has been proven for use in knowledge discovery and business intelligence in numerous markets worldwide – financial services, telecom, e-business, and online directories. Visionary knowledge management systems have fortunately begun to address these challenges as well, by incorporating enterprise search for finding results that contain not only a smattering of key words, but also discuss relevant concepts associated with the search query. The broader problem of integrating structured data with unstructured documents needs to address data cleansing and metadata management, and a combination of ETL tools and linguistic-based document processing using an enterprise search platform provide a unique solution for transforming data to actionable insight.

About FAST

FAST is the leading developer of enterprise search technologies and solutions that are behind the scenes at the world's best known companies with the most demanding search problems. FAST's solutions are installed in more than 3500 locations.

FAST is headquartered in Oslo, Norway and Needham, Massachusetts and is publicly traded under the ticker symbol 'FAST' on the Oslo Stock Exchange. The FAST Group operates globally with presence in Europe, North America, the Asia/Pacific region, South America, the Middle East and Africa. For further information about FAST, please visit www.fastsearch.com.

For any feedback or questions related to this paper, please contact us at feedback@fastsearch.com.

FAST™
www.fastsearch.com
info@fastsearch.com

Regional Headquarters

The Americas
+1 781 304 2400

Europe, Middle East & Africa (EMEA)
+47 23 01 12 00

Japan
+81 3 5511 4343

Asia Pacific
+612 9929 7725

FWP.004.01.121206