

# Bayesian Inference

Dynamic Summarization

Language Stemming

Search



Concept Searching

Personalization and Alerting

# EVOLVENT KNOWLEDGE MANAGEMENT DISCOVERY TOOLSET

## *Executive Summary*

In an environment where getting the right information to the right people at the right time is critical to an organization's success, business leaders are constantly seeking ways to enhance the effectiveness of their decision making processes. Whether it is the collection and interpretation of raw data into information or the development and implementation of creative solutions to collaboration problems, harvesting knowledge from unstructured information contained in repositories across a wide range of business operating units is the key that opens the door to an organization's intellectual capital.

In the intelligence world, adversaries strive to gain advantage over each other. Through the use of intelligence, surveillance and reconnaissance (ISR) and technical innovations, intelligence agencies are able to convey environmental awareness and information capabilities to decision makers at every level of command. Many times agencies find their missions aligned to related goals, and as a result, they end up supporting many of the same customers. When this occurs, these agencies may engage themselves in partnerships to enhance their limited capabilities.

While the idea of leveraging capabilities to gain superiority is a good goal, it is also a formidable and sometimes unattainable one. Establishing and maintaining access for intelligence community end-users to unique/segmented ISR databases, systems and secure information communications platforms results in limited end-user access and requires extensively

trained administrative and technical resources. Since the content of unstructured information is critically linked to nearly every decision making process across a wide range of business operations, organizations must be able to efficiently process their unstructured information.

Embedded within every deliberate and crisis action planning process is a decision matrix that is fueled by unstructured information. This unstructured information presents itself in many forms such as Word or PDF documents, PowerPoint presentations, or HTML and XML formatted document types. Knowledge Management (KM) solutions that fail to efficiently manage unstructured information are similar to high performance vehicles that have been filled with low octane fuel—they fail to reach their potential. Incomplete information discovery and delivery oftentimes leads to decisions that would not have been made had the decision maker been provided with a comprehensive view of a situation. On the other hand, KM solutions that offer simultaneous high recall and high precision, exceptionally high rates of information sorting/classification, language independence and scalability are like high-octane fuel in that same vehicle. Both the driver and the decision maker effectively leverage all of their capabilities.

A leader in helping organizations maximize their return on intellectual and information resources, Evolvent is the "high-octane" fuel in the KM integration business. The successful employment of unique solutions that automate manual processes used to

transform unstructured information into "actionable" knowledge, enhance risk communication to leaders and improve both information flow and collaboration are what sets Evolvent apart from those who talk about KM and those who deliver KM.

## *Introduction*

Connecting leaders with relevant and timely information in an efficient manner has always provided a competitive advantage over one's competition. In order to gain this advantage in today's business environment, leaders must implement processes and toolsets that effectively maximize the return on intellectual and information resources through the re-use of existing knowledge within their respective "enterprise."

While subject matter experts, surveillance systems and robust database analysis are crucial to converting raw data into information; it is the transformation of relevant information into an actionable context and its subsequent delivery to the appropriate decision maker that provides a leader with their advantage. The enabler of this advantage is the toolset that not only automates the collection, indexing, categorization and classification of unstructured information, but also extracts and pushes relevant information to selected decision makers through a KM platform.

Evolvent KM solutions help clients develop and integrate resources for intellectual capital management. A web-based delivery model developed to virtually share, collaborate, distribute and exchange knowledge among members of an organization,

Evolver's Knowledge Exchange (Kx) serves as an intelligent portal for content management, workflow, e-learning and collaboration. Building strategic plans that reduce costs and enhance collaboration is demonstrated by Evolver's leveraging of know-how gained by experience across a variety of vertical markets. When it comes to managing and leveraging the benefits of unstructured information, Evolver's leading edge knowledge discovery toolset embedded within the Kx provides business leaders with search and retrieval technologies that are simple to deploy, easy to integrate and adhere to current and emerging standards.

Evolver's unique combination of Knowledge Discovery technologies provides:

- Probabilistic Latent Semantic Indexing
- Relevance ranking based on the Probabilistic Model (Bayesian Inference)
- Dynamic summarization
- Concept identification based on Shannon's Information Theory
- Cross platform compatibility via Web Services
- All Application Programming Interfaces (APIs) based on XML
- Transparent access to system internals including the statistical profile of terms
- True relevance ranking of compound (i.e. multi-word) items

What does this mean to the user?

- High Recall and High Precision
- Contextualization; finding hidden relationships between documents within the enterprise
- Exceptionally high sorting rates (200,000 documents per hour)
- Language independence
- Scalability

Most KM platforms are robust document management systems. They are heavily dependent on subject matter experts (SMEs) who upload and organize information within their respective communities. This is both time-consuming and subjective, resulting in a stove-piped, static presentation of information. Retrieval of information is dependent on active searches conducted by end-users. This can be hit or miss and is compounded by a lack of integration between existing KM platforms and document management systems. Often times little or no connectivity exists to other valuable data sources.

Evolver KM solutions implemented in various business sectors effectively leverage existing technologies while driving higher utilization of the organizations' information resources. Although successful KM initiatives have a multi-phased approach, a significant return on investment (ROI) can be obtained in relatively short order. This is demonstrated through the rapid improvement in support of information discovery and communication roles. Specifically, success for this endeavor is measured by Evolver's ability to collect, index, categorize and objectively classify an organization's unstructured information.

In almost every organization, unstructured information is categorized and indexed by various individuals using "subjective" factors in a manner that is very time-consuming. Evolver's KM solution involves using "objective" criteria (established and accepted taxonomies) to automate the classification function. Unstructured information is automatically indexed and relationally organized according to Organizational, Functional (multi-disciplinary skillsets) and Product Line taxonomies. New information is collected, indexed, categorized and classified daily. Individual users can set up a profile that enables automatic notification any time relevant information is added.

The second discriminating component of Evolver's KM solution involves transforming the way an organization gets information to the end-user. While existing document management systems provide a repository where people can go for information, most do not push relevant information to the end-user. Instead of a hide-and-seek game where end-users have to visit a variety of private and public domains to obtain unstructured information, Evolver's KM solutions push content automatically from geographically separated information repositories and related sites that are relevant to what the end-user is working on.

In summary, Evolver KM solutions bring a variety of supplies and tools that are focused on providing support to a decision maker in any industry. It presents an integrated approach that automates the time-consuming process of collecting, indexing, categorizing and classifying unstructured enterprise information and results in the rapid delivery of actionable knowledge to the right person at the right time.

### *Kx Knowledge Discovery Toolset Foundation*

#### **BAYESIAN INFERENCE**

Thomas Bayes was an eighteenth century mathematician who devised a theory for conditional probability:

$$P(A/B) = \frac{P(B/A) P(A)}{P(B)}$$

Conditional probability is the probability of some event given that some other event has already occurred. In the above equation, the left hand term P(A/B) is known as the posterior probability or the probability of some event A occurring given that event B has occurred is equal to the probability of event B occurring given that event A has occurred, multiplied by the probability of event A occurring and dividend by the

probability of event B occurring.

The Probabilistic Model interprets Bayes' Theorem in an Information Retrieval (IR) context where the probability that certain query terms are better differentiators between relevant and non-relevant documents than other query terms is evaluated given implicit or explicit relevance feedback.

weights and document weights and subsequently performed extensive evaluations on relevance feedback techniques using standard document collections. In 1994, Robertson introduced an extended model that was no longer based on a binary independence model, and this work has strongly influenced the design of Evolvent's Knowledge Discovery Toolset.

Probabilistic Model. This model not only allows initial relevance ranking to be more accurate, but it also provides a mechanism for iterative searching based on relevance feedback.

#### **PROBABILISTIC LATENT SEMANTIC INDEXING**

Probabilistic Latent Semantic Indexing (PLSI) is the ability to locate documents that are relevant to the user's

**WHEN IT COMES TO MANAGING AND LEVERAGING THE BENEFITS OF UNSTRUCTURED INFORMATION, EVOLVENT'S LEADING EDGE KNOWLEDGE DISCOVERY TOOLSET EMBEDDED WITHIN THE KX PROVIDES BUSINESS LEADERS WITH SEARCH AND RETRIEVAL TECHNOLOGIES THAT ARE SIMPLE TO DEPLOY, EASY TO INTEGRATE AND ADHERE TO CURRENT AND EMERGING STANDARDS.**



#### **PROBABILISTIC MODEL**

The Probabilistic Model was pioneered at Cambridge University during the 1970's and 1980's. The model is an application of Bayes' Theorem and defines a system for weighting individual query terms and documents based on:

- The frequency of terms across the document collection (wcf)
- The frequency of terms within a given document (wdf)
- Normalized document length (ndl)
- Explicit or implicit feedback on document relevance

In 1976, Professor Stephen Robertson and Karen Sparck Jones devised a formula for computing term

#### **Why is the Probabilistic Model superior to traditional free text systems?**

Traditional free text systems are based on simple keywords and Boolean logic (primarily the AND, OR and NOT operators). While this technique is very precise, it does fall down when the number of documents retrieved is too large to examine exhaustively. In this case, the ability to rank documents, with the most important ones at the top of the list, is of paramount importance. Over time traditional systems have introduced various ways to rank results, but this is not based on a sophisticated model of term profiles across the collection of indexed documents and tend to rely too heavily on a within document frequency (wdf) analysis. The statistical model of term frequency across the document collection is unique to the

query even if they do not contain any of the words in the user's query text. It is also about the ability to ignore documents that do contain words from the user's query, but which are not relevant.

Probabilistic Latent Semantic Indexing (PLSI) is achieved by:

- Relevance ranking the documents matched by the initial query
- Extracting the distinguishing concepts from the most relevant documents
- Expanding the query to include selected related concepts

The inclusion of related concepts can be done explicitly (user decides) or implicitly where related concepts

are included automatically based on an understanding of the application area and/or user personalization.

Imagine searching for “portable computer” and finding documents that were about “aptops”, “the Toshiba Tecra” and “notebooks” but where some of the retrieved documents do not contain any words from the original query—that’s Latent Semantic Indexing.

## RELEVANCE FEEDBACK

Traditional IR systems provide a static mechanism to index documents and service retrieval requests. Relevance feedback is used to describe dynamic mechanisms that allow the retrievals to be tuned over time based on explicit or implicit feedback from the user(s). An example of implicit feedback would be where a user identifies individual documents that are relevant to their query. An example of implicit feedback would be where the system monitors the user’s activity to see what documents they examine, how long they spend looking at individual documents, what documents they author or perhaps a common pattern to their retrieval activity.

The Probabilistic Model allows this type of explicit or implicit feedback to be injected into the retrieval process so that the weightings applied are modified or tuned automatically to suit a particular user’s requirements.

## CONCEPT SEARCHING COMPARED TO SIMPLE KEYWORDS SEARCHING

A Probabilistic implementation that worked on the basis that words appear in documents independently from other words would provide a reasonable level of accuracy. However, if the implementation understands that the co-location of words is relevant and should form part of the weighting process then a significant improvement in the relevance ranking can be achieved.

For example, consider the following query:

**“Dangerous dog attacks baby”**

A human would interpret this phrase as being about a wild animal attacking an infant. However, a simple IR system that assumes that words appear independently from each other would assume that any document containing the phrase:

**“Dangerous virus attacks baby dog”**

would be 100 percent relevant to the above query on the basis that it contains all of the words. Most humans would disagree.

Evolvent’s Kx uses Shannon’s Information Theory to compute the incremental value of compound terms based on an analysis of the probability of the joint occurrence.

## SHANNON’S INFORMATION THEORY

Claude Shannon, a scientist working at Bells Labs, published his Information Theory in 1948 and this had an immediate and lasting impact on data communication technology. Shannon demonstrated that the value of a piece of information is proportional to its probability and the entropy of a joint event is given by:

$$H(x,y) = - \sum_{i,j} p(i,j) \log p(i,j)$$

Evolvent’s Kx interprets this in an IR context to compute the incremental value of a two-word term over its singleton components. Higher order compound terms are evaluated using their lower order compound components.

It is no coincidence that the majority of compound terms are in fact proper nouns, noun phrases and verb phrases, and it is these sentence fragments that convey the key concepts in most text.

However, the concepts are identified without any linguistic analysis and so the toolset works with any vocabulary and is language independent. The mathematical approach works because Shannon’s theory can be applied to any human language communication.

The ability of an IR system to identify clusters of words that identify specific concepts represents a major advancement over systems that fail to do this.

## LANGUAGE STEMMING

Often a user will type in a query with one form of a word but would like to match other forms of what is essentially the same word.

In 1980, Dr Martin Porter, a member of the team working on a Probabilistic Model at Cambridge University, developed a suffix-stripping algorithm that has been very widely adopted for normalizing words in IR systems.

Using Porter’s algorithm the following words can be matched:

“dangerous” with “danger”; “dangers” and “dangerous”

“attacks” with “attack”; “attacks”, “attacker”, “attackers” and “attacking”

“baby” with “baby” and “babies”

In addition, with our fuzzy stemmer the following words can also be matched:

“misspelt” with “mispelt”

“commission” with “commision”, “comission”, “commissioning” and “comisioned”

“accommodate” with “accomodate” and “acomodation”

Evolvent’s Kx uses language stemming as part of its concept matching process, although individual words and phrases may be left un-stemmed by enclosing with double quotes.

This means that by default, stemming broadens the matching process but where a particular word should be interpreted verbatim, it can be easily excluded from the stemming process.

**SUMMARIZATION** When a document is retrieved we normally need to display an extract from the document as an aid to the user when reviewing the returned document set. Most systems will display a static summary that is the same regardless of the user's query. Evolvent's Kx can display static summaries. However, it can also apply a modified weighting system to identify short extracts that are most relevant to the user's query. The number, length and relevance threshold for these extracts are all-configurable. The extracts will normally comprise whole sentences or short paragraphs.

#### **PERSONALIZATION AND ALERTING**

Sometimes users would like to be kept informed about a particular topic and notified when new documents arrive that are relevant to their interests. Evolvent's Kx can be used to keep users updated based on their individual profiles and will typically send an email message when new content has been added to the index.

With Evolvent's Kx Agents, the system becomes proactive, pushing content to users and eliminates the need to repeat the same searches periodically just to see what is new.

#### **SUPPORTED PLATFORMS**

The current version of Evolvent's Kx Knowledge Discovery Tools Server is available as a .NET Web Service. This means that it can be deployed on any platform that supports Microsoft.NET and may be called from any platform that supports Web Services. Therefore, it is easy for an application developer using any J2EE development environment (e.g. IBM Web Sphere) to take the Web Services Definition

Language (WSDL) file and begin making Kx API calls. The Kx Index Server is implemented as a suite of Windows programs. Sample applications are available today written in C# (for .NET), ASP (for COM+) and Java/JSP (for J2EE). A native J2EE implementation of the Query Server is also planned. The major advantage of the J2EE implementation, which has an identical API to the .NET version, will be the ability to host the Query Server on Unix platforms.

#### **CAN I CALL Evolvent's Kx FROM AN ASP/COM+ APPLICATION?**

New application development on the Microsoft platform is rapidly moving to .NET and this environment make interfacing to Web Services very simple. However, many excellent products have been developed for the ASP/COM+ environment and migrating these to .NET would be a major undertaking. Fortunately, Microsoft has provided the SOAP Toolkit for ASP/COM+ developers and using this it is fairly straightforward to call Web Services running under .NET (or J2EE).

#### **WHAT TYPES OF DOCUMENTS CAN I STORE?**

Evolvent's Kx has the following collectors:

- HTTP collector—for spidering web pages
- File collector—for documents located on file systems
- XML collector—for custom document types

Evolvent's Kx has native file conversation facilities for the following document types:

- All HTML and XML formats
- Microsoft Word and Rich Text Formats
- Adobe Portable Document Format (PDF)

- Corel WordPerfect
- PowerPoint
- Any other files in text format (e.g. TXT, CSV, etc)

In addition, an application developer can pass custom documentation types via the Evolvent's Kx XML collector.

#### **WHY IS A SQL DATABASE REQUIRED?**

The Evolvent Kx stores its probabilistic index in a proprietary database. However, the Kx uses a SQL database to manage the queue of documents to be indexed. The SQL database contains all information necessary to perform indexing, such as the individual filenames and URLs, access criteria, re-indexing frequency, inclusions and exclusions, etc. The SQL database may also be used to store any application specific meta-data.

#### *Benefits of the Evolvent Kx Approach*

#### **HIGH RECALL AND HIGH PRECISION**

Recall is a measure of how many of the documents that are relevant get found, with high recall indicating that most of the relevant documents are found. Precision is a measure of how many documents in the returned set are relevant, with high precision indicating that most of the documents returned are relevant. The Evolvent Kx offers "high recall and high precision". Others offer only "high recall or high precision."

#### **CLASSIFICATION AND SUPPORT FOR TAXONOMIES**

The Evolvent's Kx module can be used to classify documents into any predefined categories based on a small number of descriptors. Once classified the documents can then be applied to a corporate taxonomy and used for browsing the database or as a filter when running ad hoc queries. Evolvent Kx can classify around 200,000 documents per hour.

# EVOLVENT KM CONCEPT MODELS (PILOTS) PROVIDE A RAPID, MEASURABLE AND LOW-RISK ROI.



## SUPPORTED LANGUAGES

The Evolvent Kx can index any text in the Roman alphabet including full support for diacritics. The use of diacritics within documents or queries is entirely optional so that fitchée will match with fitché and vice versa. All information is exchanged and managed internally; using UTF-8 and so support for non-roman alphabets (e.g. Kanji or Arabic) should be possible in the future.

The following languages are automatically detected and processed:

- Danish
- Dutch
- English
- Finnish
- French
- German
- Italian
- Norwegian
- Portuguese
- Spanish
- Swedish
- Welsh

## SCALABILITY

The designers of the Evolvent Kx have many years experience in implementing proprietary file systems and custom databases. In particular the database format has been designed to allow concurrent indexing at full speed while allowing simultaneous access for retrievals. This concurrency has been achieved in part by reducing the amount of file restructuring typically found in competitive systems, which are often based on B-tree structures. The selected design tends to produce an index database a little larger than some alternatives but with faster retrieval. In general, the Evolvent Kx will produce an index database whose size is directly proportional to the volume of documents under index (i.e. 10GB of documents will typically produce an index database of 10GB). The proprietary database format used by the Evolvent Kx has been designed to provide optimum performance and concurrency.

For testing and development the entire system can be installed on a single computer. For live implementations the Query Server, Index Server and the Web Application would normally be distributed. A multi-server configuration will be capable of indexing about two million pages per day while simultaneously providing retrieval to hundreds of concurrent

users. For very large implementations, multiple Query Servers could be configured with shared access from a pool of application servers.

## Conclusion

Evolvent KM Concept Models (pilots) provide a rapid, measurable and low-risk ROI. A KM solution that layers over and leverages multiple document management systems, Evolvent's KM solutions lack the exorbitant integration costs traditionally found in other KM platforms and are the defining factor in determining who in business is successful and who in business is successful FIRST.