

H5 Platform

TECHNICAL WHITE PAPER



Table of Contents

Introduction	3
Technology Overview	4
A Novel Approach to Knowledge Architecture	
Advantages of H5	5
High Accuracy	
Transportability	
Concise Representation	
Scalability	
Vocabulary Independence	
Cross-lingual Capability	
Applications of H5	6
Automatic Categorization	
Content-Based Personalization	
Secure Distribution and Sale of Digital Content	
Contextual Marketing	
Expertise Discovery & Automation	
Distributed Information Retrieval	
Intellectual Foundations	8
Summary	9
Appendix: Limitations of Conventional Approaches ...	10
Bayesian Inference Model	
Vector Space Model (VSM)	
Latent Semantic Indexing (LSI)	
Semantic Networks	

H5 Platform

TECHNICAL WHITE PAPER

Introduction

In this white paper, we present the **H5 Platform**, a novel technology that aims to establish a **universal standard for managing information**, automatically measuring the subject matter content or *aboutness* of any piece of text. H5 maps text into a language-free **Subject Matter Framework (SMF)** which spans the full range of human endeavor. Using this proprietary SMF, H5 algorithms calculate a precise measurement of everything that a text is about, quantifying its relevance across all areas of human knowledge. Any text item, from a single word to an entire corpus, is assessed using this same standardized metric.

H5's content measurements are expressed in the form of an aboutness "bar code." Once bar coded, text items with similar subject matter can be matched with a high degree of accuracy, despite differences in language, vocabulary, register, writing style or syntax. Like a Universal Product Code for unstructured information, H5 bar codes not only identify everything that text is about, they enable every applications in which such data is commercially crucial: from automatic categorization, content-based personalization, expertise discovery and automation, the secure distribution and sale of digital content, and contextual marketing; to massively distributed search and cross-lingual Information Retrieval.

H5 is fundamentally unlike competing technologies, including Bayesian inference networks, semantic networks, latent semantic indexing, and vector space models. What sets H5 apart—and, more profoundly, argues that H5 bar codes are poised to become the *lingua franca* of information management—is that this technology

alone satisfies every logical condition necessary for achieving universal currency. First, H5 bar codes deliver a fine-grained reproduction of everything text is about without representing the propositional content of text. The bar codes cannot be inverted to the text itself, and thus can be used to characterize and match text with unprecedented precision and security. Second, because the SMF is corpus-independent, H5 bar codes are readily transportable and do not require indexing on a centralized server. Third, bar codes are extremely concise and can be encoded into text as metadata, remaining with the text as a guarantor of intellectual property and deliverable content. Finally, H5 is vocabulary-independent, capable of matching text items that are conceptually similar but do not contain any of the same words. Text can even be matched across languages, without the need for machine translation.

In addition to these unique attributes, H5 commands a formidable practical advantage: it scales effectively to large, heterogeneous text collections without degradations in precision or run-time performance. Moreover, since the SMF is a universal representation, it need not be customized for particular applications or customers. Rather, as flexible as the technology is scalable, H5's proprietary algorithms can be adjusted to modulate the performance of the SMF, enabling a variety of user-selected operations. Optimally concise, translatable, and scalable: these features not only capture the uniqueness of H5, they constitute the logic of its promise as a ubiquitous technology.

Technology Overview

H5 replicates two basic cognitive skills that have eluded both the Information Retrieval (IR) and Artificial Intelligence (AI) communities: assessing relevance and recognizing similarity.

H5 measures the precise subject matter relevance of any piece of text.

The H5 Encoder maps text items into a language-free Subject Matter Framework (SMF), which quantifies the amount of each type of subject matter in the text at every level of detail. The Encoder produces bar codes that record a measure of the variety of content within the text, not just the single variation or theme that generally represents it.

H5 calculates the subject matter similarity of any two pieces of text.

Given a set of H5 bar codes as inputs, the H5 Decoder calculates the distances between the bar codes. Since bar codes quantify the entirety of what text is about, distances between bar codes correspond to the full range of subject matter similarity and dissimilarity.

A Novel Approach to Knowledge Architecture

Both the Encoder and the Decoder operate on a human-built Subject Matter Framework (SMF), which consists of hundreds of thousands of hierarchical subject matter fields that represent important distinctions within domains of human endeavor. The proprietary SMF has several highly specific characteristics that distinguish it from other forms of knowledge representation, such as ontologies, common sense reasoning formalizations, lexical hierarchies, and semantic networks. The primary advantage of the SMF is that it reproduces actual human relevance judgments rather than simply relying on statistical relationships between words. Embedded in each node of the SMF are the terms and phrases that are uniquely relevant to (and

hence highly reliable indicators of) specific subjects.

The SMF cannot be machine generated.

Machines can analyze the frequency of terms in text, but term frequencies are not an adequate substitute for human relevance judgments. For example, analyzing texts on the Microsoft antitrust case, a statistical engine would naively infer that because “operating system” and “lawsuit” frequently co-occur, these terms have some necessary relationship. Like a person, H5 Technologies H5 recognizes that the two terms belong to different fields of knowledge, one to Computers and the other to Law, and that only in texts like these are they found together.

H5 discerns implicit relevance relationships.

When people read text, they draw on a wealth of tacit knowledge to interpret even the simplest statements. Text that can be easily understood by a human reader is difficult for most text-processing systems because conventional software cannot access this store of implicit knowledge. For example, the term “cells” is relevant to the fields of Microbiological Systems, Prison Housing, and Electrical Equipment. If an IR system saw the word “cells,” it could only hazard a guess about the correct sense of the word, using thesaurus-based probabilities. By contrast, accessing the wealth of the implicit relevance relationships embedded in the SMF, H5 pinpoints the subject matter fields present in the document, and can deduce the correct sense of the term on the basis of context. Furthermore, these same unstated relevance relationships can be leveraged to match documents on the basis of increasing specificity or generality. For example, H5 could match a document referring to the concept “object oriented” or to the more general category of “programming languages” or to the more specific concepts of “Java” or “C++.”

Advantages of H5

Products based on H5 offer the following technical advantages over competing technologies:

High Accuracy

Because H5 measures differences in subject matter, or *aboutness*, instead of differences in keywords, it is free from the performance limitations of word-based techniques. Once the H5 Encoder has identified what a text item is about, the Decoder can automatically filter out text items on irrelevant subjects, ensuring a high degree of accuracy.

Transportability

Unlike statistical approaches to text processing that remain fixed to the collections they describe, H5 is corpus-independent. H5 bar codes will always be the same, regardless of the repository in which the text items are stored. As a result, text collections do not need to be indexed on a central server, and can be readily exchanged among different communities of users. New text items can be added without the need for re-indexing the entire collection, and separate collections can be readily correlated or merged without extra programming.

Concise Representation

With conventional text processing technologies, the profile representation is as large as (if not larger than) the original text. By contrast, the average H5 bar code is only 5% the size of the text it represents. The conciseness of H5 bar codes means that they can be encoded into text as metadata and rapidly exchanged across a massively distributed network.

Scalability

H5 scales effectively to large, heterogeneous text collections without degradation in precision or run-time performance. In addition, H5 remains efficient even when run on huge text collections. Thanks to the elegant hierarchical structure of the SMF, text can be encoded and decoded with great

speed. The algorithms are easily parallelizable, and dispense with computational complexities.

Vocabulary Independence

H5 solves the vocabulary problem. Unlike competing technologies, H5 can match two text items whether or not the items contain the same words. The richness of representation in the SMF eliminates dependence on simplistic keyword or thesaurus-based approaches. The SMF has a total vocabulary of more than 400,000 terms: ten times that of a highly literate academic or professional and 20 times that of the average person. In addition, each field in the SMF employs all the terms or phrases that a person could use to express a particular concept. For example, the field of Labor Shortages incorporates the terms “labor crisis,” “shortage of workers,” “employee shortage,” “employment crises,” and so on.

Cross-lingual Capability

The globalization of commerce and information makes the development of effective cross-lingual technologies increasingly important. The H5 Decoder is capable of accepting a user’s request in his or her native language and then seamlessly searching, retrieving, relevance ranking, and displaying text items written in a variety of foreign languages. The SMF can be embedded with differentiating terms from any language, from English to Mandarin Chinese. H5 is therefore of special significance to organizations whose knowledge requirements include foreign language sources.

Applications of H5

Once encoded into a text item, H5 bar codes enable a multitude of applications, from the desktop, to the extended enterprise and the Internet.

Categorization

Efficient categorization of documents is a major challenge for enterprises that handle massive quantities of emails, reports, memos, etc., on a daily basis. These organizations require automated systems that can rapidly route documents to specific in-house users, organize them into pre-specified categories, or even send requested portions of documents to the appropriate downstream affiliates.

H5's patent-pending algorithms for assessing subject matter relevance allow for accurate, automated categorization using only a small set of sample text items. H5 scales effectively and is capable of identifying extremely specialized subject matter distinctions. By contrast, Bayesian algorithms—the prevailing approach to text categorization—can neither scale nor make specialized subject matter distinctions. (See **Appendix**.) For example, Bayesian systems have difficulty distinguishing among documents about different operating systems, such as UNIX, Linux, Windows, and Windows NT. In addition, Bayesian categorization systems require extensive training using a large set of sample text, and subsequent changing of the categories usually requires retraining the entire system.

Content-Based Personalization

The H5 platform provides for the personalized delivery or retrieval of text based on User Interest Profiles. H5 creates highly specific User Interest Profiles by aggregating the bar codes for all of the text items a user reads or writes. By maintaining a set of User Interest Profiles, online publishers,

corporations, and news organizations can offer a range of personalized services. As users' needs change, their Interest Profiles adapt without the need for manual intervention. Moreover, the user may also fine-tune weights assigned to the various subjects in his or her interest profile. The system requires no retraining whatsoever, and categories automatically conform to a user's evolving interests.

Secure Distribution and Sale of Digital Content

One of the current barriers to the sale of online digital content is the poor performance of text retrieval technologies. Before customers will spend money on digital content, they must be sure that the content retrieved for their request will actually be relevant to their interests. The challenge for content distributors—such as online publishers—is to reveal enough about the content to entice customers, while safeguarding valuable intellectual property. H5 gives customers confidence that the digital content they buy will actually match their interests, and does so without compromising the security of the underlying content.

Contextual Marketing

The greatest challenge facing content sites on the Web is to monetize the content they publish, by providing users with contextually relevant advertisements and merchandise offerings. Fortunately, online users always indicate what interests them in the same way: **by reading content**. Yet current technologies lack the sophistication to exploit text as an index of user interests. Current technologies have therefore failed to deliver effective automated contextual marketing solutions.

Capturing the nuanced relevance of content with remarkable precision, H5 Technologies H5 now succeeds in delivering dynamic, automated,

targeted advertising and product recommendations.

Expertise Discovery & Automation

Categorizing, sharing, retrieving and utilizing knowledge assets is a challenge of daunting proportions for today's organizations. Current technologies have so far failed to meet the challenge of harnessing corporate knowledge and expertise, causing an estimated \$31 billion annual knowledge-deficit for Fortune 500 companies alone.

H5's unique approach to Expertise Automation unlocks information about the knowledge, priorities, and project focus of each individual in the organization. H5 continuously profiles employee knowledge and interests without human intervention, connecting users to directly relevant information, resources, and human experts throughout the extended enterprise.

Distributed Information Retrieval

With the availability of storage and network bandwidth, H5 now makes it possible to link widely distributed, heterogeneous text collections together to form meta-collections of unprecedented scope. H5 is ideally suited for distributed search for several reasons. First, the vocabulary for specifying an information request is completely independent of the text items in the collection. Second, H5 bar codes are extremely concise and can readily be transmitted between computers even with low bandwidth. Third, the bar codes consist of a universal, quantitative subject matter representation that is powerful enough to distinguish text items in a single collection, yet generalizable enough to match text items across collections. Wherever text items may be located or hosted, H5 bar codes enable the immediate identification and retrieval of all items relevant to a user's request.

Example: H5 Bar Codes Add Value to News

The news industry, like most industries, is moving from a product-based business model to a service model, due in part to the widespread availability of free news on the Internet. In the face of an increasingly commoditized information marketplace, news organizations have stepped up efforts to differentiate their services by developing new techniques for personalizing their content. The recent adoption of standardized metadata formats such as the News Industry Text Format (NITF) and News Markup Language (NewsML) reflects the desire of the news industry to make the delivery of news faster, more efficient and more personal. NewsML and NITF enable news services to encode a variety of metadata describing what a news item is about, to whom it may be of interest, its general importance and so on.

While standardized XML encoding schemes like NewsML and NITF provide a necessary framework for news personalization, these formats are only as effective as the information that they encode. The proprietary tagging schemes currently used for identifying the subject matter of news items are not well suited for personalization because they are not precise enough to capture the specific interests of individuals. Precise distinctions are crucial, however, for representing the interests of highly-educated financial specialists who comprise a much sought-after audience for financial news services.

(continued on page 8)

Intellectual Foundations

Pioneered by company co-founder and Chief Scientist Dr. H. Joel Jeffrey, the H5 Platform owes its origins to the energizing synthesis of Dr. Jeffrey's principal areas of expertise: Computer Science and Descriptive Psychology (DP).

Providing a unique conceptual framework for generating precise, comprehensive, and nuanced articulations of human behavior, DP has informed groundbreaking research in a wide variety of specialized disciplines: clinical psychology, social psychology, organizations and organizational intervention, software engineering, and AI.

Chief exponent of DP's practical application to expert systems and document retrieval, Dr. Jeffrey has developed a theoretical stance quite unlike those of traditional approaches to text processing. Whereas conventional IR generally assumes too simple an equivalence between a text's words and purported underlying "concepts," Dr. Jeffrey's innovative work reconceptualizes the business of classification: categorizing the myriad human activities and cultural practices that particularize the sense of each and every word in any text. Hence, the H5 platform effectively sidesteps the pitfalls of using statistically derived clusters of words as surrogates for subject matter fields. From the perspective of DP, the project of "reading between the lines" looks misdirected, for—read rightly—"the lines" themselves are the bearers of *aboutness*.

As early as 1976, prototypes of Dr. Jeffrey's novel information processing system proved successful, achieving unmatched precision and recall within a restricted domain. Dr. Jeffrey faced two major challenges, however: there was no obvious way for the model to scale beyond a restricted domain and the model relied on excessively tedious data gathering techniques, requiring the manual encoding of millions of human relevance judgments.

(continued from page 7)

H5 bar codes render proprietary tagging schemes obsolete, allowing news services to match User Interest Profiles and content in highly specific ways. H5 bar codes are easily integrated with both NewsML and NITF. In addition to increased precision, H5 offers several unique advantages that enable news services to optimize their news content personalization:

H5 bar codes make it easy for end users to create customized User Interest Profiles. Users may either identify specific subjects that interest them or select sample news items that match their interests.

As user interests change, User Interest Profiles automatically zero-in on increasingly relevant information. The weights assigned to specific subjects in the User Interest Profiles can also be easily adjusted based on individual preferences.

User Interest Profiles are language-independent. Given a User Interest Profile, one can choose to receive relevant documents in a variety of languages.

Dr. Jeffrey met these challenges by engineering a hierarchical knowledge base, the first SMF, which radically reduced the need for manually encoded relevance judgments. By replicating the structure of domains of human endeavor and developing algorithms for measuring H5 bar codes, Jeffrey allowed the H5 platform to scale across an unlimited number of domains.

Summary

The H5 Platform represents a new paradigm for the processing of unstructured text. By measuring text against a universal, language-free Subject Matter Framework (SMF), H5 is the first technology capable of automatically assessing everything that text is about. From single words to entire archives, H5 submits any piece of text to the same universal standard, making it possible to match an unlimited number of texts about similar subjects, despite differences in language, keywords, and syntax.

H5 supersedes previous technologies, offering high precision, transportability, concise representation, scalability, vocabulary independence, and cross-lingual processing. Together, these advantages satisfy every logical necessity for achieving technological ubiquity, positioning H5 bar codes to become the *lingua franca* of information management. As such, H5 naturally enables a panoramic range of applications: from automatic categorization, content-based personalization, expertise automation, the secure distribution and sale of digital content, and contextual marketing; to massively distributed search and cross-lingual Information Retrieval.

The solidity of its theoretical grounding; the suppleness of its computational flexibility; the representational breadth and depth of its Subject Matter Framework; the conciseness, semantic density, and rapid transportability of its bar codes—all confirm that the **H5 Platform is singularly qualified to serve as the universal standard for information management.**

Appendix: Limitations of Conventional Approaches

Until now, all Information Retrieval (IR) technologies have relied on substituting the assessment of term similarity for the measurement of subject matter similarity. But the realities of human language usage stymie these attempts.

Bayesian Inference Model

Probabilistic models for Information Retrieval have been in use for at least forty years, but companies such as Autonomy first exploited the commercial potential of probabilistic systems in the early 1990s. Bayesian systems use statistical inference to find apparent correlations among clusters of words in a text collection.

Inherent limitations of the Bayesian approach preclude it from becoming a universal standard. Bayesian systems are incapable of making fine-grained distinctions in subject matter, and do not scale well to large, heterogeneous document collections due to computational complexity. In addition, the Bayesian model is strictly corpus dependent, rendering impossible the transport and exchange of documents.

The key limitation of a Bayesian system is that it first must learn the distribution of term probabilities specific to a given collection of documents and then is hampered by the unreliability of the very probabilities on which its inferences are based. In all but the most restricted corporate or academic settings, the text collection used to derive the statistics simply cannot approximate the full range of real world subject matter concepts. As a result, when a document contains a combination of terms at variance with the expected distribution of terms, Bayesian systems perform poorly. While an automated software tool that fails to embrace unanticipated requests is merely annoying, a knowledge management system incapable of appropriately routing mission-critical documents that contain

some previously unforeseen combination of words might jeopardize an entire enterprise.

Vector Space Model (VSM)

Developed by Gerald Salton in the 1960's, VSM associates a set of terms with each document or request. VSM retrieves documents by finding the best match between the request terms and the document terms, often comparing the frequency of terms in the document to the expected frequency given the collection of documents as a whole. The weakness of VSM is that all matching is done on the basis of superficial word similarity. The use of a thesaurus for query expansion can only partially mitigate this weakness. As a result, **VSM cannot discriminate among documents that do not share any terms, and therefore rarely achieves the desired level of performance.**

Latent Semantic Indexing (LSI)

LSI statistically analyzes the patterns of word usage across the entire document collection. Using a statistical technique known as Singular Value Decomposition, LSI derives clusters of terms in documents. These term clusters are then used to assign to each document a list of term clusters. Documents with similar term clusters are placed near each other in a term-document space, allowing documents to be matched, even though they may not share terms. As with VSM, these term clusters are sometimes referred to as "concepts." **However, many patterns and categories of human behavior are simply not reflected in word-use patterns that can be derived statistically.** As a result, the performance of LSI represents a marginal improvement over other techniques.

Semantic Networks

Semantic networks, growing out of research in knowledge representation within Artificial

Intelligence, have been used sporadically for Information Retrieval over the past forty years, with very little success. In this model, a database of common sense knowledge about the world and human interactions is organized as a vast network that mediates between queries and documents. **Not surprisingly, semantic networks do not scale beyond very restricted domains.** Not only must semantic networks be built by hand, but the complexity of the real world means that a network covering even a small portion of its intricacies is utterly impractical.

H5 Technologies, Inc.

World Headquarters

520 Third Street, Suite 17

San Francisco, CA 94107, USA

<http://www.h5technologies.com>

Worldwide Inquiries:

Phone: 415-625-6700

Fax: 415-625-6799

The H5 Platform is a trademark of H5 Technologies, Inc.

All other company and product names mentioned are used for identification purposes only, and may be trademarks of their respective owners

© 2001 H5 Technologies, Inc. All rights reserved.