Beyond Keywords: The Revolution in Search

Joe Buzzanga, MBA, MLS, MA
Senior Product Manager, IEEE

Introduction

In a digital world search is essential, pervasive, ubiquitous and ceaseless. Yet, search remains unsatisfying in many ways. Why is this the case? When we search, we’re after an answer. But today’s search engines don’t have enough intelligence to provide answers. They’re also not advanced enough to interpret naturally phrased queries and understand the ambiguities of ordinary language. And so we type in keywords and get links or lists of results that are more or less relevant depending on the type of search, our query and the competence of the search engine. There is always a further step when we have to select an item from a result set and assess it. Hopefully, the answer eventually emerges, but there are no guarantees.

The ultimate search engine would directly answer our query. Perhaps it would even anticipate our query and present the answer before we even ask. To answer us, the search engine would need to understand our query, our context and the corpus of knowledge at its disposal—in short, it would need Artificial Intelligence (AI). The problem of search ultimately resolves to a problem of AI.

A deep AI search engine remains a futuristic vision for the most part. But search technology is advancing rapidly and recent innovations are beginning to overcome the limitations of the conventional keyword query/result list model. Search is reflecting glimmers of something resembling “meaning” or “intelligence”. Meaning in search is emerging along three related dimensions. First, meaning in the sense of enormous knowledge bases that provide a foundation for supplying direct answers to queries. Second, meaning in the sense of semantic understanding automatically derived from unlabeled document collections through machine learning algorithms. Third, actionable meaning in the guise of intelligent personal assistants that provide deeply contextual search triggered through a natural, conversational interface.

The Building Blocks of Search: Classic Information Retrieval

Modern search engines are applications of the technologies developed in the field more formally known as Information Retrieval. In this paper we treat search and information retrieval as synonymous concepts. To understand the revolution in search, we need to briefly touch on the
foundation technologies in classic search or information retrieval systems. The principal information retrieval models are listed in Table 1 (adapted from Baeza-Yates and Ribeiro-Neto 2011, 60)

**Table 1: Information Retrieval Models**

<table>
<thead>
<tr>
<th>Type</th>
<th>Example</th>
<th>Result Ranking</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set Theoretic</td>
<td>Boolean</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>Algebraic</td>
<td>Vector Space, Latent Semantic Indexing</td>
<td>Term Frequency *</td>
<td>Score quantifies the probable relevance of document to query</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Inverse Document Frequency</td>
<td></td>
</tr>
<tr>
<td>Probabilistic</td>
<td>BM25, Bayesian, Language Models</td>
<td>Document score</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>PRIMARYLY based on link distribution + weighted term matching</td>
<td>Google weighs over 200 “signals” to rank a page</td>
</tr>
<tr>
<td>Graph Based</td>
<td>Page Rank</td>
<td>PRIMARYLY based on link distribution + weighted term matching</td>
<td></td>
</tr>
</tbody>
</table>

These models determine how a query is processed and how results are matched and ranked. They each have various strengths and weaknesses, a discussion of which is beyond the scope of this paper. Modern search products typically blend elements of each model. For example, the popular open search platform, Lucene, uses both vector space and Boolean.

All of these models struggle with the protean, elastic and inherently ambiguous aspects of human language, especially polysemy and synonymy. Polysemic words have different meanings but are spelled and may be pronounced the same—for example “bank” as in “river bank” and “Federal bank”. Synonyms are different words with the same or similar meaning—for example, “engine” and “motor”. Polysemy and synonymy undermine precision and recall of search results. The emerging search technologies discussed later in this paper are beginning to overcome these linguistic issues by matching queries to documents based on meaning rather than keywords.

**Reference Model for Search**

The techniques developed in these retrieval models support a search process which can be broken down into a number of simple elements as depicted in Figure 1 (adapted from Hearst 2009). The parts outside the brackets—Information need and Corpus—are external to the search engine.
The starting point for the search process is a user and an information need. The need is expressed in a query. The search engine interprets or processes the query and selects candidate responses from its index. The responses or results are ranked and returned to the user. The user evaluates the results and if satisfactory, the process stops here. Otherwise, the user may elect to reformulate the query and assess another set of results. Note that the search engine is dependent on its index for results. The index is a representation of a corpus or collection of content.

**Information need**

The information need reflects a problem or question that the user is trying to resolve. The possibilities are virtually limitless here, ranging from the prosaic and utilitarian (“should I wear a coat today?”) to more complex and abstract needs (“what position should I take on global warming?”). In the former case, a simple, fact-based answer may be what is desired (“Yes, wear a coat, it is below freezing outside”) while in the latter case the need may reflect a desire for a quick education on the subject. With the rise of the web and now smartphone apps, users expect more and more of their information needs to be satisfied through a search.

**Context**

An information need exists in a specific user context. Examples of context include the user’s search and purchase history as well as any personal preferences voluntarily submitted to the search engine. With mobile search an entirely new set of contextual information can be made available to a search engine including location in time and space, locally installed apps, bodily vital signs and indicators and more. Capturing a rich user context is an area of significant innovation and competitive differentiation in modern search engines, allowing personalization and tailoring of results that match the characteristics of the user.
Query
The user converts or adapts their information need into a query in order to search. Query input has evolved from intricate, syntactically sensitive Boolean interfaces usable only by trained searchers to looser keyword and phrase-based interfaces accessible and usable by all. Search engines do significant processing of queries, ranging from simple word stemming, to more advanced probabilistic expansion and interpretation techniques. The actual entry of the query is still largely based on textual input of keywords or phrases via physical or virtual keyboard interfaces. Query input is evolving rapidly though, and leading web search engines and emerging intelligent personal assistants are making progress in speech and conversational query interfaces.

The formulation of a good query can make the difference between a successful and unsuccessful search. Although modern search engines have made great strides in interpreting queries and in handling quasi-normal phrases and questions the query input is still far from being a natural or intuitive way to ask a question. Conversational and natural language based search interfaces are an active area of research and innovation in search technology.

Query Taxonomy
There have been many attempts to create a systematic categorization of queries (Hearst 2009). An influential classification was developed by Broder (2002) based on web log analyses. Although limited to the web, it still provides a useful basis for discussion of search queries in general. Queries fall three types: Informational, Navigational, and Transactional.

- **Informational**: The intent is to acquire some information assumed to be present on one or more web pages. These may also be called “discovery” queries
- **Navigational**: The immediate intent is to reach a particular site. A user seeking a specific site, document or person, somewhat like using a directory; may also be called a “known item search”
- **Transactional**: The intent is to perform some web-mediated activity (purchase an item, download a file, etc.)

Departing from Broder, we refine the informational category to two sub-types: Informational/Answer and Informational/Discovery.

- **Informational/Answer**: Query may be posed as an actual question and the intent is to get a direct answer (“Who invented the transistor”)
- **Informational/Discovery**: Query likely posed as a phrase and the intent is to learn or become knowledgeable about a subject ( “wireless communications technology”)

Index
The index is a data structure that represents and stores items drawn from a content collection or corpus. Most modern search engines use an “inverted index” which lists all the
Designing an efficient, scalable, resilient and robust index is a source of competitive advantage and differentiation for search engines, especially in the case of consumer web search engines which must index petabytes of data and handle hundreds of thousands of queries per second with millisecond response time.

**Corpus**

The corpus is the collection of material indexed by the search engine. The main corpus for search engines like Google and Bing is the web, although they are beginning to index information stored in mobile apps. Specialized search engines limit themselves to particular collections of content. An example of a specialized search engine is Scopus, a classic abstracting and indexing database covering a corpus of peer reviewed scientific and scholarly literature. Patent search products like Derwent World Patent Index are another example.

**Results and Ranking**

The search engine processes the query and looks for matches against the index. Results are selected, ranked and presented to the user. The most widely used ranking technique in traditional search engines was pioneered by Spark Jones and Salton in the 1970s (Sanderson and Croft, 2012). The intuition was that terms that occur frequently in a document, but infrequently in the overall collection, are the best discriminators for relevance. The concept is mathematically expressed in the formula: \( \text{TF} \times \text{IDF} \). Page and Brin proposed Page Rank (Page and Brin, 1998) as a more useful mechanism for the web.

**A Taxonomy of Search**

The discussion so far has centered on search in general. However there are many kinds of search engines and they differ along various dimensions. Some are fee-based and focus on scientific, technical or other scholarly material. Others are freely available to the general public and index the web and more. Still others aren’t search engines proper, but nonetheless offer rich discovery and navigation features. A high level classification and description of search engines is provided in Table 2.
Table 2: Search Engine: High Level Taxonomy

<table>
<thead>
<tr>
<th>Search Category</th>
<th>Examples</th>
<th>Overview</th>
</tr>
</thead>
</table>
| General Web              | Google, Bing, Baidu, Yahoo, Yandex           | • **Audience:** universal, consumer-oriented  
• **Corpus:** enormous ([60 trillion pages](#)) per Google, uncontrolled and includes text, images, video, and audio  
• **Index:** petabytes+  
• **Results:** deliver links but rapidly evolving to deliver answers and facts  
• **Query input:** largely textual via keyboard, but supports spoken queries as well |
| Technical/Scientific     | Scopus, Summon, Derwent World Patent Index    | • **Audience:** Specialized—technical, scientific users  
• **Corpus:** controlled and numbers in the millions of documents; mainly text but increasingly capturing figures, images, video and audio  
• **Index:** terabytes  
• **Results:** metadata records including abstracts  
• **Query input:** textual via keyboard |
| Intelligent Personal Assistant | Siri, Cortana, Google Now                 | • **Audience:** universal, consumer-oriented  
• **Corpus:** hybrid—web+structured knowledge bases+personal apps  
• **Results:** direct answers+back-off to web links  
• **Query input:** spoken natural language |
| Answer Engines           | Wolfram Alpha, IBM Watson                    | • **Audience:** specialized  
• **Corpus:** mainly structured and authoritative knowledge bases  
• **Results:** direct answers  
• **Query input:** natural language questions, mathematical formulas |

**Meaning (Facts) from Unstructured Information**

**Semantic Web Vision**

Tim Berners-Lee, James Hendler and Ora Lassila presented a vision for the next generation of the web in the May 2001 article “The Semantic Web” (Berners-Lee et al. 2001). In the article, Berners-Lee and his co-authors vivified the idea of a semantic web through a detailed
use case that imagined an intelligent web software agent capable of checking calendars, making appointments, finding trusted persons and places and more. The “semantic web agent” does all of this autonomously, drawing inferences on behalf of its human user. Berners-Lee tells us that all of this can be achieved without futuristic, sci-fi-like AI. It becomes possible through the encoding of meaning or semantics into web pages by their authors. But it’s not just web pages. Interestingly, the vision also included intelligent physical entities like home appliances that adjust their settings in concert with the needs of the household residents.

Inherent in the semantic web vision are three core concepts.

1. A web imbued with meaning expressed using Resource Description Framework (RDF) “triples” (subject/predicate/object)
2. Ontologies identifying the things that exist, their definitions and their relations
3. Software agents (“semantic web agents”) capable of inference and autonomous action

Progress?

Fourteen years later, where do we stand? In one sense, we are not close to realizing the vision. There aren’t any software agents roaming an open, semantically enriched web, drawing inferences from reliable factual information and completing tasks for users. On the other hand, bits and pieces of the vision are blossoming although they are taking shape in ways unanticipated back in 2001. For example:

- The major search engines are increasingly extracting meaning from the web, leveraging semantically tagged pages and large structured knowledge bases.
- Intelligent personal assistants like Siri, Google Now, and Cortana have emerged which resemble the predicted “semantic web agent”. However, so far they lack inferential ability. In addition, they navigate a hybrid digital space composed of the open web and “closed” smartphone apps. This significantly diverges from the more completely open vision sketched by Berners-Lee.

From Hypertext to HyperKnowledge: Unleashing the Knowledge Base

A large part of human knowledge exists in the form of factual statements or simply, facts. What if a search engine could learn facts about the world? If harnessed correctly, it could use the facts to more intelligently serve up information to searchers. But what are facts?

One general class of facts can be expressed as things or entities with properties. For example, the statement “Obama is president” is a fact. It states a relationship between and entity (Obama) and a property (president). It can also be viewed as a class-instance relation with “Obama” an instance of type “president”. Facts like these are being collected into enormous

---

1 We know this concept today as the Internet of Things.
knowledge bases such as DBpedia, NELL and others. The major internet companies are assembling their own versions, including Google, Microsoft, Facebook, Amazon, and IBM.

<table>
<thead>
<tr>
<th>Name</th>
<th>Entities</th>
<th>Facts</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>DBpedia</td>
<td>4.6M</td>
<td>583M</td>
<td>Wikipedia</td>
</tr>
<tr>
<td>YAGO</td>
<td>10M</td>
<td>120M</td>
<td>Wikipedia, WordNet, GeoNames</td>
</tr>
<tr>
<td>NELL</td>
<td>5.2M</td>
<td>50M Candidates, 2.4M Confident</td>
<td>Machine learning, crawls open web</td>
</tr>
<tr>
<td>Knowledge Graph (Google)</td>
<td>500M</td>
<td>3.5B</td>
<td>Wikipedia, Freebase, and other sources</td>
</tr>
<tr>
<td>Knowledge Vault (Google)</td>
<td>45M</td>
<td>271M</td>
<td>Machine learning, crawls open web</td>
</tr>
</tbody>
</table>

Introducing these knowledge bases into the search process can help to disambiguate queries and deliver answers to users. They are, in a sense, “shadow” indices which amplify and enrich standard inverted indices.
Commercial Web Search Examples

Google introduced its knowledge base in 2012 under the name “Knowledge Graph”. The Knowledge Graph largely relies on crowd sourced and manually curated information. While it has an estimated 3.5 billion facts, the manual and crowd sourced methods impose limits on its scale and completeness. In 2014 Google launched a related project called “Knowledge Vault” which uses machine learning techniques to probabilistically learn and fuse facts from any source on the web. This may permit the Knowledge Vault to vastly expand Google’s knowledge of the world. Knowledge Vault uses 16 different extraction systems to capture facts from webpages (Dong et al., 2015). In addition, it is able to calculate a probabilistic confidence score for facts. Knowledge Vault has 45 million entities, 1.6 billion triples and 271 million confident (0.9 probability or better) facts (Dong et al., 2014).

Google is using this factual data in at least three ways:

1. To better understand query intent
2. To return factual structured answers to queries
3. To support a Question/Answer interaction with Google Now, its intelligent personal assistant

It may also use Knowledge Vault confidence scores as an additional signal to rank the quality and relevance of web pages (Dong et al., 2015).

Microsoft launched a knowledge base under the name “Satori” in March 2013. It has been somewhat elusive about the number of entities included, but trade press reports peg the number at 400 million (Gallagher, 2012). Like Google, it captures facts and returns structured answers to certain types of queries in a user interface (UI) element it calls “Snapshot”.

Figures 4 and 5 are examples of so called “knowledge panels”, illustrating how the two companies are using their knowledge bases to more intelligently handle queries.
Both systems capably handle the query, which is directly posed as a question rather than a keyword phrase. Because they “know” that the Eiffel Tower is a building with the property...
“height”, they can deliver a direct answer to the query. They can also provide a profile of the Eiffel Tower with additional relevant information. Google also presents fact-based answers in “structured snippets”, introduced in Q3 2014.

Google, Bing, Yahoo and Yandex are actively encouraging web developers to include structured data in their web pages using the Schema.org mark-up vocabulary. As this process takes hold in the developer community we can expect to see more structured, factual answers returned in search results.

Knowledge bases are helping search engines serve users who enter Information/Answer style queries noted earlier in this document. Despite the advanced technology operating behind the scenes to deliver these knowledge panels or snippets, the results are still fairly rudimentary. The search engine returns basic facts about certain classes of entities in response to relatively simple queries. The knowledge bases, though populated with millions of entities, are still capturing only fragments of what exists in the world, mainly certain well known classes of people, places and things. And they are a long way off from the more ambitious goal of autonomous logical inference. But these technologies are still nascent. They are developing rapidly and there is massive commercial investment in talent and compute infrastructure to take them to the next level.

Meaning Quantified

Machine Learning—Neural Networks

Building knowledge bases from facts is one avenue to a more intelligent search. But not all queries can be satisfied by a simple factual answer and not all documents contain facts. And even for those that do, the subject/predicate/object triples may not adequately exhaust or expose the essential meaning of the document. Older keyword systems, of course, make no pretension of capturing meaning or concepts. Are there other technologies that could intelligently process or derive a deeper conceptual meaning from documents and queries?

Machine learning, a sub-discipline of AI, is beginning to have an impact here. In machine learning, computers apply algorithms to big data to detect patterns and make predictions. When used with search applications, machine learning can identify the essential concepts in documents and queries, whether explicitly expressed or not. The machine learning algorithms do not rely on a priori taxonomies, ontologies or knowledge bases and can be applied to unstructured data like textual documents. A specific branch of machine learning known as “deep learning” or “deep neural networks” has gained a lot of attention in recent years.

Deep neural network systems are loosely modeled on the neural structure of the human neocortex and try to enable computers to learn somewhat as humans do. Although the origins of
deep neural networks go back to the 1980s, the technology had failed to deliver on its promise until recently. The missing elements were massive data sets to feed the network and sheer computational power. Today we have both.

Recent progress has been impressive. Google, for example, is using deep learning in 47 products launched since 2013 including Android speech recognition, Street View and Ad placement (Dean, 2015). Perhaps the most striking success in recent months was demonstrated by a deep learning program that taught itself to beat the classic video game Breakout (Kumaran and Hassabis, 2015).

While deep neural networks have had the most success in speech recognition, computer vision, and other non-textual tasks, there are commercial applications emerging in information retrieval. A company called Cortical.io has created an API using the technology that can be applied in many natural language processing use cases, including information retrieval. Their neural network solution is based on open source learning algorithms from Numenta, a Silicon Valley start-up.

Another company, IP.com, has developed a patent search product called InnovationQ based on its own “shallow” or artificial neural network technology. The system is based on a so-called “unsupervised” learning model, meaning that labelled, pre-categorized documents are not required. It simply ingests the native or raw documents and automatically learns their inherent conceptual meaning. The system matches queries to documents based on concepts, not keywords which allows it to overcome the problems of polysemy and synonymy.

In this approach the query is treated just like a document, meaning that better results are gained by using more lengthy queries (although the system will “back off” to a more traditional keyword search if the query is just a couple of words). This can be a real benefit, since it frees the searcher from having to figure out the best keywords and query structure. Instead, a chunk of text can be directly entered in the search box.

Figure 6 shows how the system handles a query that is input as a block of text. The query is a narrative description of an invention. This is a realistic use case in a patent search scenario, where an inventor may submit a description of the invention to a searcher to conduct a prior art search. Here is the query:

In this invention, we propose to utilize analog behaviors of off-the-shelf Flash memory to enable hardware-based security functions in a wide range of electronic devices without requiring custom hardware. More specifically, we show that a standard Flash memory interface can be used to generate true random numbers from quantum and thermal noises and to produce device fingerprints based on manufacturing variations. The techniques can be applied to any floating-

---

2 The differences between an artificial neural network (ANN) and a deep neural network are beyond the scope of this discussion.
gate non-volatile memory in general, and does not require any hardware modifications to today's Flash memory chips, allowing them to be widely deployed.

Figure 6: Visualizing Results from Semantic Search

The figure shows a concept map, one of the ways in which result sets can be viewed. Each dot is a document. The large terms crystallize the key concepts from related document clusters.

Machine Learning—Topic Models

Topic models are probabilistic algorithms that discover the latent conceptual or topical structure of a document collection. They do not require prior annotation or labelling of the documents and they make no use of exogenous ontologies or knowledge bases. Similarly, they ignore the syntactic structure of sentences and even the sequence of words, instead representing a document as a “bag of words”. The topic modelling approach to concept identification is purely statistical. It assumes that a document is generated from a set of topics and that topics are probability distributions over words. By grouping words based on probability, these systems overcome polysemy and synonymy—words naturally fall into the proper topic, regardless of their lexical expression.

Figure 7 illustrates a topic model used to describe a specific article in the journal Science (Blei 2012). The topics—Genetics, Evolution, Disease, Computers—represent the underlying conceptual content of the article. They are generated roughly as follows. A probabilistic algorithm (Latent Dirichlet Allocation) is applied to a document collection of 17,000 articles
from Science. This results in a set of 100 topics that pervade the collection. Then the model calculates the relative mixture of these topics in each article. The topics in Figure 7 are those that are most densely represented in the article.

<table>
<thead>
<tr>
<th>“Genetics”</th>
<th>“Evolution”</th>
<th>“Disease”</th>
<th>“Computers”</th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>evolution</td>
<td>disease</td>
<td>computer</td>
</tr>
<tr>
<td>genome</td>
<td>evolutionary</td>
<td>host</td>
<td>models</td>
</tr>
<tr>
<td>dna</td>
<td>species</td>
<td>bacteria</td>
<td>information</td>
</tr>
<tr>
<td>genetic</td>
<td>organisms</td>
<td>diseases</td>
<td>data</td>
</tr>
<tr>
<td>genes</td>
<td>life</td>
<td>resistance</td>
<td>computers</td>
</tr>
<tr>
<td>sequence</td>
<td>origin</td>
<td>bacterial</td>
<td>system</td>
</tr>
<tr>
<td>gene</td>
<td>biology</td>
<td>new</td>
<td>network</td>
</tr>
<tr>
<td>molecular</td>
<td>groups</td>
<td>strains</td>
<td>systems</td>
</tr>
<tr>
<td>sequencing</td>
<td>phylogenetic</td>
<td>control</td>
<td>model</td>
</tr>
<tr>
<td>map</td>
<td>living</td>
<td>infections</td>
<td>parallel</td>
</tr>
<tr>
<td>information</td>
<td>diversity</td>
<td>malaria</td>
<td>methods</td>
</tr>
<tr>
<td>genetics</td>
<td>group</td>
<td>parasite</td>
<td>networks</td>
</tr>
<tr>
<td>mapping</td>
<td>new</td>
<td>parasites</td>
<td>software</td>
</tr>
<tr>
<td>project</td>
<td>two</td>
<td>united</td>
<td>new</td>
</tr>
<tr>
<td>sequences</td>
<td>common</td>
<td>tuberculosis</td>
<td>simulations</td>
</tr>
</tbody>
</table>

**Figure 7: Topic Model for a paper based on 17,000 articles from the journal, Science (Source: Blei 2012, 79)**

In information retrieval, topic models have been used for document modeling, topic extraction, and concept browsing. They have also been explored for use in automated ontology creation (Wei and Barnaghi, 2010) as well as a range of non-textual, pattern recognition applications. Toolkits for experimenting with topic models are easily found on the web. One of the most popular is an open source package called [MALLET](https://mallet.cs.umass.edu/).

**Actionable Meaning: Intelligent Personal Assistants**

Intelligent personal assistants are exemplified in products like Siri, Google Now and Cortana. The industry and trade press hasn’t quite settled on a name for the product category though, underscoring the relative immaturity of this new class of software. Besides “intelligent personal assistants”, they are also referred to as “personal digital assistants”, “virtual assistants”, or “intelligent automated assistants”. Apple used the term “knowledge navigator” in a prescient 1987 video. Today it uses “intelligent personal assistant”, and we will follow this convention in our discussion.

The name “intelligent personal assistant” reflects its three key defining features:
- **Intelligent**—conversational user interface that simulates an understanding of natural language and directly responds to the user with answers or actions (inferential capability is on the horizon)
- **Personal**—unlike search, these products adapt to the user in new and profound ways
- **Assistant**—aim to offload routine tasks (schedule a meeting, make an airline reservation...)

Intelligent personal assistants aren’t strictly search or information retrieval tools, although that is part of what they do. They aim to go much farther than search. Generally speaking, their intent is to provide answers and perform actions in a personalized, highly contextually aware manner. They even want to anticipate the user’s needs and present individually tailored, context relevant information before the user even thinks to ask. These products are well on their way to displacing certain types of searching behavior and as they develop it’s possible that they may completely superecede search as we know it today.

Table 4 gives a sample of some intelligent personal assistants. Some are well known, others are still in development.

<table>
<thead>
<tr>
<th>Vendor</th>
<th>Product</th>
<th>Initial Release</th>
<th>Scope</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apple</td>
<td>Siri</td>
<td>October, 2011</td>
<td>General</td>
</tr>
<tr>
<td>Google</td>
<td>Google Now</td>
<td>July, 2012</td>
<td>General</td>
</tr>
<tr>
<td>Microsoft</td>
<td>Cortana</td>
<td>April, 2014</td>
<td>General</td>
</tr>
<tr>
<td>Samsung</td>
<td>S Voice</td>
<td>May, 2012</td>
<td>General</td>
</tr>
<tr>
<td>Nuance</td>
<td>Ask Nina</td>
<td>August, 2012</td>
<td>Task specific – Customer care</td>
</tr>
<tr>
<td>x.ai</td>
<td>x.ai</td>
<td>Restricted beta</td>
<td>Task specific – Scheduling assistant</td>
</tr>
<tr>
<td>Viv</td>
<td>Viv</td>
<td>In development</td>
<td>General</td>
</tr>
</tbody>
</table>

These products are still in an early, formative stage and while there are no objective empirical evaluations or systematic performance benchmarks available, anecdotal and informal tests in the trade press generally report mixed results. One recent report tested the question answering ability of Google Now, Siri and Cortana, based on over 3,000 queries. Google Now returned a direct, correct answer to 51% of the queries, while Siri managed to answer 15% correctly and Cortana only 8%. The test did not evaluate their performance in carrying out actions or other “personal assistant” related tasks.
Results like these may seem underwhelming, but when one considers the complexity and sophistication of the underlying technology integration and innovation embedded in these products, their performance is impressive. And most importantly, this is only the start. Their improvement is inevitable.

**Intelligent Personal Assistant--Technology**

Although work on intelligent personal assistants goes back decades, it wasn’t until Apple introduced Siri that the category crystallized as a real technology with mass market potential. For this reason, we turn to Siri here as the exemplar for the product category.

Siri provides a variety of services some of which are illustrated in Figure 9. Many of the services rely on third party sources such as Yelp, Bing, Wolfram Alpha, Wikipedia, Rotten Tomatoes and others. This information is combined with personal profile information drawn from a user’s contacts, calendar, music library, email etc., The final ingredients are temporal, spatial, location and other sensor derived data. All of this is combined with automatic speech recognition (ASR) and text to speech (TTS) technology allowing the user to interact with the system in a natural, conversational manner.

A glimpse into one of the many Siri-related patents provides further insight into the underlying technology.
Siri’s cognitive core is an “Active Ontology”\(^3\). This is a unifying infrastructure that manages the various components of the system including domain models, natural language processing, dialog management, memory and services interfaces and task flow models. The Active Ontology combines these functional blocks with an execution environment that lets it accomplish tasks that fall within the scope of its domain models.

**Sentient Search**

The fact that this technology has co-evolved with mobile computing is no coincidence. The smartphone is the perfect device to fuse and simulate digital intelligence. It is inherently networked, it has massive processing power, and most of all, it is virtually sentient. The camera (vision), microphone (hearing) and embedded sensors allow the phone to partially mimic the human perceptual apparatus. Even more sensory abilities are coming with the force sensing “taptic” engine in the Apple Watch.

In the future, this sentient functionality will be combined with inferential reasoning to provide a profoundly personal, contextual software agent that will adjust search in real time.

---

\(^3\) The Active Ontology concept was developed at SRI in connection with research on an intelligent personal assistant that was the precursor to Siri.
based on your location, who you spoke to last, who is around you, etc. Answering traditional search queries will be provided almost as an afterthought. It is difficult to guess if devices or software with these capabilities will be realized in the near future or if it will be much further out. But it is clear that this is the vision the vendors are working towards.

**Conclusions**

The major web search vendors have progressed far beyond keywords and page rank. They are using knowledge bases, machine learning, natural language processing and other technologies to deliver answers, interpret natural language queries and understand the semantics of documents and queries. Intelligent personal assistants are making progress in conversational interfaces, answering queries, discerning context and accomplishing simple tasks. It is logical to suppose that many of these technologies will begin to be integrated into the premium technical and scientific search products used in libraries. The other possibility is that these products may be further marginalized as the general purpose web search engines become intelligent enough to effectively subsume them.

**Implications for Information Professionals**

Information professionals have a tremendous opportunity to help their organizations realize value from emerging machine learning tools and knowledge bases. To do this, they need to become conversant with the field of data science. This doesn’t mean they need to be programmers and data scientists themselves, but they need to be comfortable with the key concepts in data science and the latest commercially available products. Whether in a leadership role, or in partnership with IT, the information professional can provide needed expertise in evaluating and selecting new information management and discovery solutions that incorporate machine learning and knowledge bases.

These new technologies hold the promise of converting almost any unstructured textual data (email, customer support logs, internal documents etc.) into a knowledge asset. Information professionals should seek out opportunities to play a role in these applications, whether advising on ontologies and taxonomies or providing guidance on discovery and browse interfaces. This means taking a broader view of information, one not restricted to traditional artifacts like books and journals.

Really proactive information professionals will experiment with machine learning and knowledge bases to enhance discovery of their content collections. They will seek out creative ways to bring data science expertise into their departments and will work with vendors to apply natural language interfaces, semantics and meaning to their internal search infrastructure.
References


http://googleblog.blogspot.co.uk/2012/05/introducing-knowledge-graph-things-not.html.


