



Factiva Synapse White Paper KMS101
Taxonomies, Ontologies, Thesauri and Authority Files:

The Key to Better Information Retrieval

Hidden Information

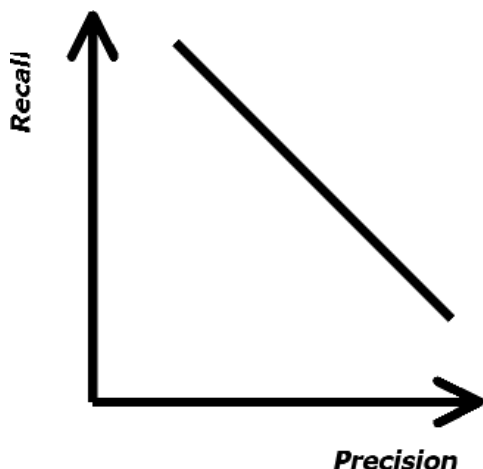
Few quotes so efficiently describe both the achievement and yet the failing of the information age at the beginning of the twenty-first century. We have terra-bytes of digitized information and global networks connecting millions of computers, but finding specific information is more than ever like looking for a needle in a haystack.

One of the major reasons why information retrieval is problematic is because language is elastic: one word can have many meanings, and one concept can be expressed by many different words. Human beings can understand words better than machines because our brains store a vast and complex web of interrelationships in which no word is an island.

Text-based searching has failed to provide a satisfactory method for interrogating large data sets and will fail increasingly as the sets of data we need to access continue their exponential growth.

The problems of text-based searching have been known in information science for decades as the precision-recall tradeoff.

Empirical studies have shown that precision will decline as recall increases and vice-versa; this tradeoff is inherent in the very nature of text-based searching.



Recall Quotient relevant items retrieved / universe of relevant items

Precision Quotient relevant items retrieved / total items returned

In other words, if you need to ensure you see all the relevant information then you will inevitably have to wade through a greater mass of irrelevant data—whatever you do to improve the comprehensiveness of your search will diminish the proportional purity of the results.

“Information networks straddle the world. Nothing remains concealed.

But the sheer volume of information dissolves the information. We are unable to take it all in.”

– Günther Grass

The Human Paradigm

The human brain has approximately 100 billion neurons, but it is not this capacity that gives us our noble reason. Our infinite faculties derive from the way the brain is wired through a complex network of interconnections called synapses. The number of ways this network can interconnect is not measured in billions; it is a number greater than all the fundamental particles - electrons, protons, neutrons, etc. - in the entire universe. Nature revealing the infinite within the finite.

Interconnectedness is the key to the way the human brain works and also the key to solving the precision-recall dilemma in information systems.

In the human mind words are not isolated islands like they are in machines. Every thought, word and image is intricately connected to other related words and concepts through many subtle relationships. If we want machines to be able to understand our requests for information, and to respond with comprehensive and relevant results, then we need to give them a knowledge-base that is structured the way our own brains work.

The Machine Correlative

The word ‘syndetic’ means to bind together or to connect; in information science it is applied to the cross-referencing of vocabulary terms to represent a variety of conceptual links.

Standard methodologies have evolved for building relationships between concepts in information systems using controlled vocabularies. Broadly speaking these include hierarchical, equivalency and associative relationships.

Ultimately, the goal of a controlled vocabulary is to represent each discrete real-world object or unique abstract concept by one and only one unambiguous indexing term, and then to cross-reference these indexing terms to represent the rich interconnections inherent in their conceptual or real-world correlative. Homographs (words with more than one meaning) need to be disambiguated, and synonyms (the same meaning represented by more than one word) need to be mapped to one hub term known as the preferred term. Concepts are arranged in one or more hierarchies to represent the various categorical organizations by which the domain may be viewed. Concepts are also linked together by free association with other related concepts to further enrich the syndetic network. These types of interconnections are discussed in more detail in the following three sections.

The Elasticity of Language

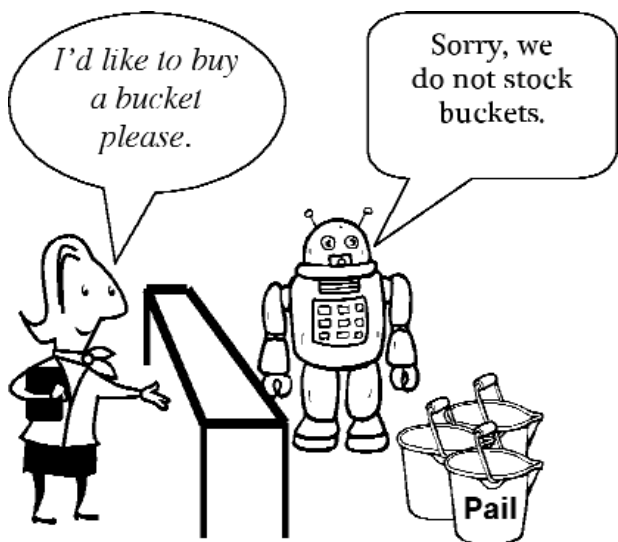
Dr. Peter Roget, author of the first thesaurus of the English language, described a phenomenon he called 'the elasticity of language'. Put simply: one word can refer to several different things (homographs), and one thing can be referred to by several different words (synonyms) .

For example, the word 'mercury' is a homograph. It could be used to refer to the Roman god of communication, a planet in our solar system, a chemical element, or the trade name of a line of automobiles.

Something you might use to carry water in could variously be referred to as a 'bucket' or a 'pail', these are examples of synonyms.

If these variances seem innocuous it is because our brains instantly and subconsciously translate synonyms on-the-fly and differentiate homographs by an internal knowledge-base of subtle contextual associations. But the variances are far from innocuous when it comes to information retrieval systems.

Imagine going into a hardware store run by a computer and asking to buy a bucket...



The robot-shopkeeper may adamantly respond that he does not sell buckets, despite the fact that there are a dozen pails piled up in the back room.

This is the most insidious consequence of the elasticity of language in text-based systems: relevant information often remains hidden. In a controlled vocabulary, equivalency relationships are used to map synonyms together and one of the synonyms is picked to be the preferred term for indexing purposes.

Less insidious but just as frustrating is the homograph problem where, for example, an online search for astrological articles on Mercury may result in the user wading through thousands (or on the web, millions) of irrelevant articles on gods, cars and chemicals. Refining the search by building a Boolean construct such as 'mercury' and 'planet' will improve the precision of the search, but comprehensiveness will necessarily suffer – some relevant results will now be excluded because they do not contain the word 'planet'. In a controlled

vocabulary, homographs are disambiguated by the use of parenthetical modifiers, e.g., 'Mercury (Planet)'; in a taxonomic structure disambiguation might be achieved by the context of the term's placement in a particular hierarchy.

Equivalency relationships may be used for other kinds of links besides connecting synonyms. They can be used for mapping spelling variants, regional and multilingual equivalents, organization-specific preferred terminology, acronyms and abbreviations.

Hierarchical Perspectives

Hierarchical relationships are used to classify concepts; they provide a top-down organization of ideas that present an ontology (categorical view of reality) for any given domain of knowledge. They are often presented as taxonomies, ontologies or classifications. In a hierarchy, tree-structures are presented such that each lower level branch has some differentiated attributes while possessing the qualities of its parent branch. Differentiation can be of three types: generic, whole-part or instance.

Generic relationships indicate that each narrower term is a kind of its parent term, e.g., cacti are a kind of succulent plant.

Plants

Succulent plants

Cacti

Whole-part relationships indicate that each narrower term is a part of (component of) its parent term, e.g., spark plugs are a part of the ignition system of an automobile.

Automotive electrical system

Ignition system

Spark plugs

Instance relationships indicate that the narrower term is an instance of its parent term. The parent term is usually a common noun category whereas the narrower term may be a proper noun, e.g., Beringer Founders' Estate is an instance of the Merlot subcategory of red wines.

Red wines

Merlot

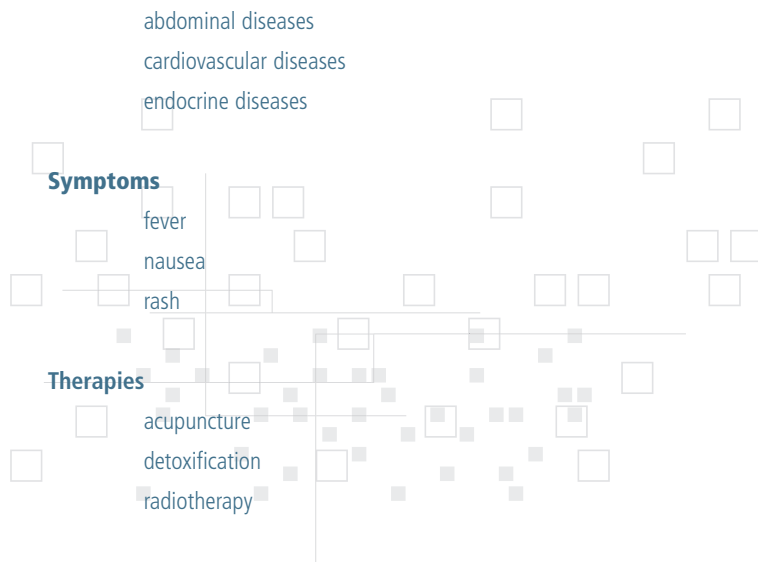
Beringer Founders' Estate

Often terms may logically belong in more than one hierarchy, e.g. musical instruments might branch down paths for stringed instruments and percussion instruments; since pianos are a kind of both stringed and percussion instruments, the term pianos would therefore require placement in both structures. This is called polyhierarchy.

One of the problems of hierarchies is that there can be as many different ways of ordering a hierarchy as there are users of the information. The structure depends on the perspective of each viewer.

For example, when building a medical ontology one could identify several major views, each of which lends itself to hierarchical organization:

Diseases

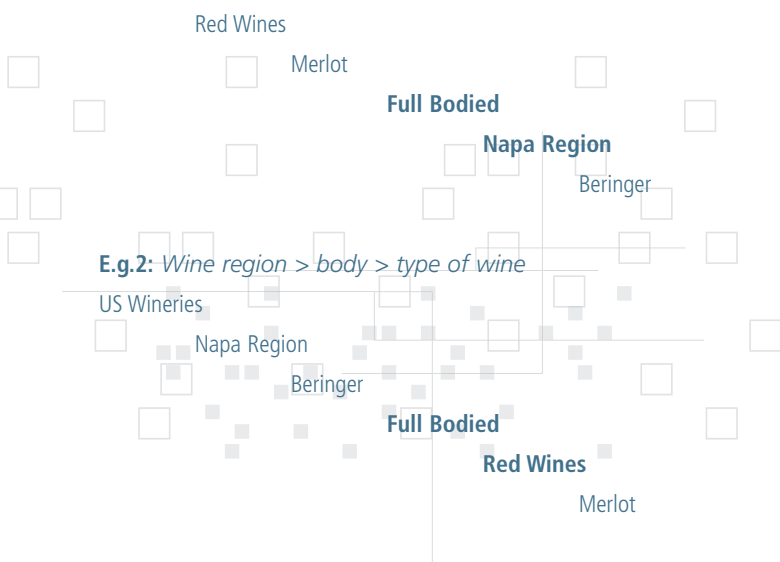


It would not be possible to build a single hierarchy that could structure all the permutations of these concepts in a way that would satisfy the variety of user-perspectives. In such instances it may be desirable to create a set of independent hierarchies representing each major facet of the domain of knowledge.

Facets present new challenges in user interface design as multiple hierarchies need to be navigated and dynamically recombined. One possible solution may be described using a metaphor of parallel universes and wormholes. In this model a graphical process allows the user to jump through a succession of separate ontologies or facets with a single query. The wormhole method is unique in that it presents the results as if they are a continuous hierarchical pathway, even though the actual pathway does not exist in the knowledge-base; it is dynamically constructed by the user. In the following two examples the arrow-bullets indicate where the user has wormholed from one facet to another; in both cases the results display like a continuous hierarchy.

E.g.1: *Type of wine > body > wine region*

Wines



E.g.2: *Wine region > body > type of wine*

Associative Thinking

Beyond hierarchical and equivalency relationships lies the even richer and more subtle world of associative relationships. These relate concepts and objects together based upon a great variety of contextual associations. For example, temperature is related to thermometers, harvesting is related to crops, and death is related to bereavement and may even be related to its antonym life.

The links in all of the above examples represent very important associations, but none of them could be expressed through the hierarchical structures of a taxonomy nor through equivalency relationships. These associative relationships are the hardest to define, but they provide the richest and most subtle connections between concepts bringing machines closer to the kind of knowledge-base that human beings take for granted.

Electronic Thesauri

An electronic thesaurus represents the ultimate machine-intelligent knowledge-base for storing the myriad forms of interrelationships that can exist between concepts.

Unlike taxonomies, which only store hierarchical associations, an electronic thesaurus can combine the categorical order found in taxonomies, equivalency and mapping relationships as well as more intuitive associative links.

National and international standards exist to guide the construction of such thesauri, e.g., ANSI/NISO Z39.19 and ISO 2788 and 5984.

Named-Entity Control

The discussion so far has concentrated on conceptual vocabularies, but controlled vocabulary techniques may also be usefully applied to other types of vocabulary, such as personal names, organizational names, geographic entities and other proper-noun lists.

Named-entity authority control is a branch of information science that utilizes the principles of conceptual vocabulary control. The goals of named entity authority control are:

1. to disambiguate different people, organizations or other entities with similar or identical names;
2. to bind together all the variant forms of names that they use;
3. to organize entities hierarchically where appropriate;
4. to display other types of association between entities.

With such a knowledge-base, machines can pull together a more complete picture about people, organizations and other entities regardless of the many variant forms of names that may be used within different information sources, and regardless of the particular form known to the inquirer. Special methods need to be employed to allow specific variant forms to be used for particular associations (e.g., Samuel Clemens was the author of certain books published under his legal name and other books under his pen name Mark Twain). Despite allowing these variant-differentiated associations, it is

required to be able to cluster all associations together under a common hub for the entity.

Personal Names

Pseudonyms, aliases, nicknames, married / maiden names, AKAs, working names, short names, long names, code names, legalized names, pen names (noms des plumes), stage-names and sobriquets all serve to confuse and confound the proper identification of people or organizational entities in information systems. All of the above may be examples of proper and legitimate variant forms of names. The problem can be even further compounded when individuals or organizations willfully attempt to conceal their identity.

Formal and informal relationships and associations between individuals, such as marriage, parentage, friendships and business relationships, can also be expressed using customized associative relationships.

Disambiguation of individuals with similar or the same names may need to invoke rules-based validations that reference controlled metadata elements such as birthplace, residences, birth and death dates, state and other official identification numbers, and many other data.

Organizational Names

Organizational names apply to formal entities like corporations, academic institutions, non-profit organizations and local and national government organizations and their executive branches. They may also embrace less formal groupings of people such as performance art groups and literary circles. Covert organizations may also be covered such as criminal syndicates, political groups and terrorist cells.

Like personal names, organizational names also suffer from the confusion of aliases, pseudonyms and a plethora of other variant forms. Organizations frequently change their name through the process of mergers, acquisitions and partnerships. Many organizations can be arranged with some hierarchical structure such as the corporate division / subdivision structure.

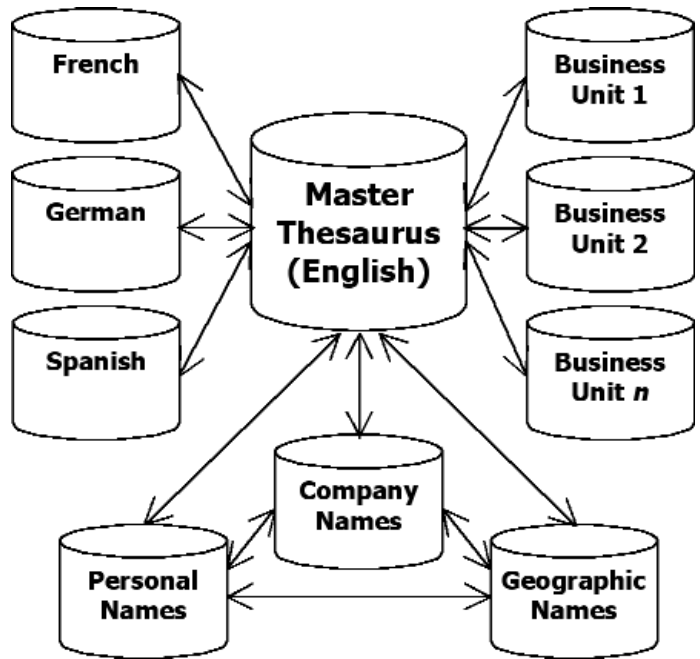
Important associations between organizations can be identified such as Company A merged with Company B, Company C customer of Company D, etc.

Personal named entities may also be associated with organizational named entities to show the membership of specific individuals in particular organizations.

Geographic Names

Geographic names apply to countries, states, regions, cities, subdivisions, territories, seas, rivers, mountains, deserts, forests, jungles, plains, and many other geographic and topographic entities. These entities also have variant forms of names, identical names are duplicated in different places and the names evolve through history with shifts in territorial boundaries.

Most geographic entities lend themselves to hierarchical structuring as well as other associative link



Enterprise Knowledge-Bases

For most large corporations and government organizations the goal is to build a master controlled vocabulary that unifies all the data sources in the enterprise. The enterprise might have some business units that use specialized, private, legacy or third-party vocabularies. By mapping these satellite vocabularies to the hub of a master vocabulary the organization can benefit from cross-searchability throughout the entire enterprise. This configuration may be called a metathesaurus.

Following is a diagram of a model enterprise-wide knowledge-base.

Multilingual meta-thesauri, private and third-party vocabularies and named-entity vocabularies are all mapped through a single syndetic network. Note: all conceptual vocabulary, including multilingual versions, map to a common hub (the master thesaurus); the master thesaurus and the named entity files are all mapped to each other.

Six Degrees of Separation

The phrase six degrees of separation was coined by John Guare in his 1991 stage play of the same name. The concept is that any one can make a connection to any other person on earth through a chain of six intermediary people; each link in the chain must be between people personally acquainted or familiar with each other. At first the proposition sounds unbelievable, but the play was inspired by empirical research undertaken by Harvard professor Stanley Milgram in 1967. Milgram's research was based on connecting two people randomly picked from anywhere in the USA; his study resulted in an average of 5.5 intermediary people.

What we are observing in the six degrees of separation phenomenon is the close proximity between nodes in complex networks, and how fast and easy it is to get from one place to another in remarkably few leaps (phone calls, mouse clicks, etc.).

Conclusion

Text-based searching necessarily fails to yield accurate and comprehensive results due to the inherent elasticity of language. The human brain manages to overcome this problem because humans have a sophisticated knowledge-base in which no word is an island. In order for machines to be able to achieve the same level of precision and recall when searching for information, they need controlled vocabularies with rich syndetic networks.

Scale is important - small and simple taxonomies are not sufficient for accessing the rapidly growing information repositories managed by large corporations and government organizations.

Syndetic networks, however, can disambiguate similar concepts and names, pull together variant forms, capture equivalencies between multilingual vocabularies, arrange concepts and named entities into multiple hierarchical groupings, and capture subtle associative relationships between concepts and named entities.

The richer the network's cross-referencing within and between these vocabularies, the more successful will be the syndetic network at answering user queries accurately, comprehensively and quickly.

