Chapter 2
Concepts

A concept is an entity of consciousness. We know a concept when we encounter one "in action", because it exceeds its stand-in descriptive label as a word, phrase or sentence. A concept might be a directly conceived or an intuited object of thought. In general, every object, issue, idea, person, process, place, etc., can generate a concept. Although concepts are an integral part of human cognition (or perhaps precisely because of this fact), their exact definition is fairly difficult. Various viewpoints and approaches to their explanation are presented in the first part of the present chapter. To be of any practical use for representing knowledge, concepts cannot appear in isolation, but must be associated with each other. Thus, in the continuation of the chapter, the most important formalisms for organizing concepts are presented along with the examples of organizations in actual applications. The main practical application of concepts is document retrieval based on the actual meaning referred to with the search phrase instead of the one based on (more or less) literal phrase matching. The last section of the chapter is dedicated to the topics related to concept-based search.

2.1 Definition

The term "concept" comes from Latin word *conceptum* ("that which is conceived"). Although the explanation of a concept has been a mainly philosophical matter ever since the ancient era, it is also a subject matter of psychology and linguistics. Throughout the centuries, many theories about the nature of concepts have been proposed, many times fundamentally contradicting each other. It all seems that when trying to explain what a "concept" is, only few things hold for certain—the most apparent definitely being the fact that the unique definition of a concept does not exist.

Despite the challenging explanation and the absence of a uniform definition, it is very often accepted that concepts are somehow connected with the process of human cognition. It is also commonly established that concepts are *abstract* and *universal*.

G. Jakus et al., *Concepts, Ontologies, and Knowledge Representation*,
SpringerBriefs in Computer Science, DOI: 10.1007/978-1-4614-7822-5_2,

The abstractness of concepts arises from the fact that concepts do not enclose the specifics of or differences between the objects to which they apply. In this sense, all objects to which a concept applies (also referred to as "extensions" of concepts) are treated as indistinguishable from the respective concept. Because concepts apply to every object in their extension in the same way, they are considered universal.

The abstract nature of concepts does, however, not prejudice the nature of objects in a concept's extension. These objects can thus be abstract or concrete, real or imaginary, atomic or composed of other objects. A concept can be anything existing in the human mind or shared through human language and behavior, for example an action, a task, a strategy or way of thinking (Gomez-Perez and Corcho 2002).

These are merely some of many viewpoints from which one can define the nature of concepts. In the continuation of this subsection, we have chosen to present the most important concept definitions grouped by the criterion whether concepts are defined explicitly or implicitly.

2.1.1 Explicit Concept Definitions

2.1.1.1 Concepts in Philosophy

The beginner of the classical theory of concepts is the Greek philosopher Aristotle. In his "theory of definition", he defined a concept by using a pair of two concepts which he named *genus* (a kind, sort or family) and *differentia* (a distinguishing characteristic). According to his theory, the concept "*human*" is, for example, defined as a "*thinking animal*". The "*animal*" corresponds to *genus* as it determines the family human belongs to and "*thinking*" corresponds to *differentia* as it distinguishes humans from other animals.

Philosophers Locke and Schopenhauer considered concepts to be abstractions of what is obtained by some sort of individual experience in the form of sensation and reflection (Locke 1690; Schopenhauer 1851). Besides the concepts abstracted from human experience (the so-called a posteriori concepts), the German philosopher Immanuel Kant also mentions another type of concepts: those that originate in the human mind (Kant 1800). Kant referred to these concepts as *pure* or *a priori* concepts or *categories*. Two examples of a priori concepts are the concepts of time and space.

Another German philosopher, Gottlob Frege, argued that a concept reflects the way in which we comprehend the world around us (Frege 1892). Frege gives an illustrative example which associates an object ("reference"), a symbol ("sign") and a concept ("sense") revealing the differences among them. The symbols *morning star* and *evening star* refer to the same object—planet Venus. However, the senses of these two terms are completely different, since the former is visible in the morning and the latter can be observed in the evening. The two symbols,

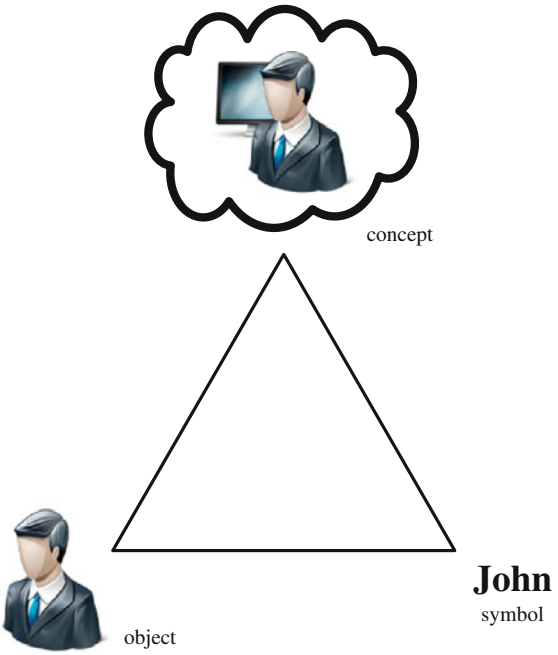

Fig. 2.1 The meaning triangle

therefore, represent two different concepts which correspond to different observations of the world, i.e. the time of observation in our particular example.

A similar explanation is given by John Sowa who considers a concept as a "*mediator that relates symbol to its object*" (Sowa 2000). Such mediation can be illustrated with the so-called meaning (or semiotic) triangle introduced by (Ogden and Richards 1923). In the lower two corners of the triangle in Fig. 2.1, there is an icon resembling a person named John and a printed symbol representing John's name. The cloud on the top represents a "neural excitation" that is induced by an object associated with the symbol representing the respective object. The "neural excitation" depicted in Fig. 2.1, for example, represents John working at his office and appears in one's mind when thinking about a person named John. This excitation, a mediator between the symbol and its object, is called a concept.

2.1.1.2 Concepts in Other Scientific Fields

In contrast to philosophy, where concepts are directly associated with the very essence of human existence and perception of the world, the treatment of concepts in other fields of science is often more pragmatic. There, the definitions are tailored to depend upon the way concepts are used in practice and are usually expressed in terms of the field's terminology.

In linguistics, a concept is most often considered as a unit of meaning formed through the abstraction of concrete words and phrases and as such corresponding

to the so-called *conceptual meaning*. A good example illustrating the linguistic treatment of concepts is WordNet—a lexical database containing interconnected English nouns, verbs, attributes and adverbs (Miller 1995; Fellbaum 1998; Princeton 2010). WordNet defines concepts (although implicitly) through sets of synonym words called *synsets*. For example, a synset consisting of synonym words *homo*, *man*, *human being* and *human* defines a concept that can be lexically expressed by any of the words in the synset. In addition, WordNet complements such treatment of concepts with an explicit definition expressed in the description of the synset "concept" ("*an abstract or general idea inferred or derived from specific instances*").

The WordNet example indicates that a concept can be represented with more than one word. On the other hand, however, individual words can represent more than one meaning as well, and can, therefore, stand for more than one concept. Another relation between words and concepts is also significant: concepts are language-independent, while words are not. When translating among different languages, the meaning, though expressed with different words, can be preserved to a large extent.

Beside the above-presented, several other readings of concepts in linguistics are identified in (Smith 2004). For example, a concept can also be understood as "*a meaning that is shared in common by the relevant terms* [*and/or*] *in minds of those who use these terms*".

In the field of engineering, a concept is related with building models of entities from reality. From such a perspective, concepts can be defined as "*creatures of the computational realm which exists … through their representations in software, in UML diagrams, XML representations, in systems of axioms*", etc. (Smith 2004).

In contrast to engineering, a concept in mathematics transcends the reality that is recognized through our emotions and intuition. The mathematician Carl Benjamin Boyer defined mathematical concepts (such as the integral or derivative) as "*well-defined abstract mental constructs*" which are "*beyond the world of sensory experience … although they may be suggested by observation of nature or intuition*" (Boyer 1959).

2.1.1.3 Concepts in Knowledge Representation

As already mentioned in the introduction, concepts are often considered as atomic elements in knowledge representations. Therefore, it is worth looking at how concepts are treated in this field.

One of the formalisms used to represent organized knowledge are concept maps. Here, concepts are defined as a "*perceived regularity in events or objects, or records of events or objects, designated by a label*" (Novak and Cañas 2008). The label for most concepts is a word or a symbol.

Description logic is a formalism for representing logic-based knowledge through concepts (classes), roles (relations) and individuals (objects). In description logic, concepts denote sets of individual objects (Baader and Nutt 2002).

Frames (described in Sect. 4.2.1.1) are knowledge representation structures influenced by the organization of human memory. In contrast to many other formalisms for representing knowledge, concepts are not considered as atomic units when represented with frames. Instead, they are treated as sets of highly structured entities which can be described with recursive structures consisting of pairs of attributes (called "slots" in frame terminology) and their values (Petersen 2007).

Closely related to frames are object-oriented languages. Object-oriented languages recognize two elemental structures: *classes* and their instances, referred to as *objects*. A class defines the properties and methods that can be used to manipulate the properties of all the objects that are instances of a particular class. The properties and methods defined in a class provide the objects with a state and behavior. As classes act as abstractions of concrete objects, they correspond to concepts while objects correspond to instantiations of concepts.

2.1.2 Implicit Concept Definitions

If concepts can be recognized, but cannot be defined exactly and consistently, how can a machine, for example, distinguish which words in a text represent concepts and which do not? Which terms carry more meaning than others? How can one make concepts recognizable, so that they can be automatically extracted from any type of texts?

A Vector Space Model (VSM) (Salton and Wong 1975) presents one possible answer for the above questions. In VSM, each document is represented as a vector with coordinates representing the frequency of the observed index terms in a document. The number of coordinates (the dimension) of each vector corresponds to the number of index terms observed in a document collection. For example, in Fig. 2.2, three different documents are represented by three two-dimensional index vectors, whereby the two dimensions correspond to two index terms observed in the documents.

A level of similarity between two documents can be measured by calculating the inner product of the corresponding index vectors or the inverse function of the angle between them (when the angle between two vectors is zero, the similarity function is at a maximum, and vice versa). The latter approach is also used in the example from Fig. 2.2.

The similarity among documents can also be measured when a new index term is assigned to a document collection. If the similarity level decreases, the newly assigned index term has a high discrimination value and is as such considered as a "good" index term. The opposite holds for a "bad" index term. Therefore, "good" index terms can be recognized as concepts, since they represent the smallest units of knowledge carrying as much meaning as possible (i.e. enough to decrease the similarity level between the documents when assigned to a document collection).

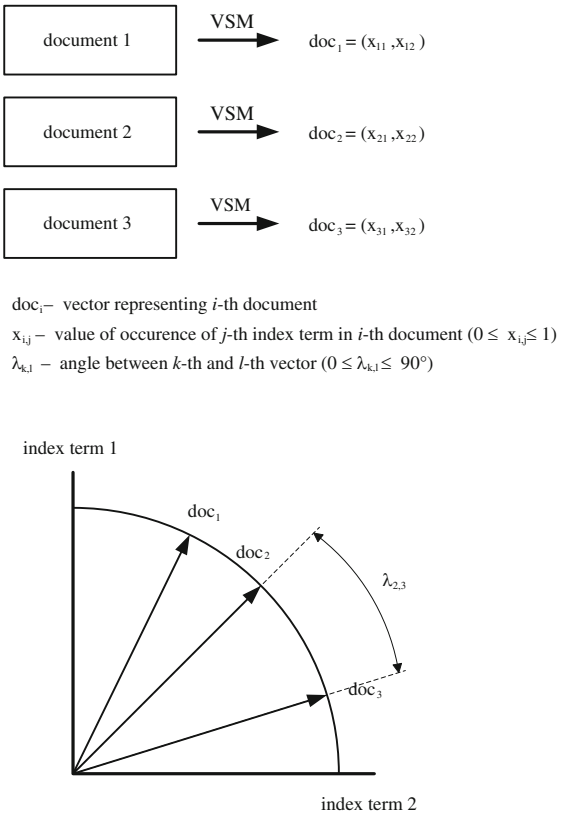

Fig. 2.2 Representing documents by two-dimensional vectors in the Vector Space

Figure 2.3 shows document representations when a "good" discriminating term is added to VSM. Before assigning the index term 3, the three document vectors reside on one plain formed by the axes of index terms 1 and 2 (Fig. 2.2). After adding the index term 3 to the document collection, a third dimension is added to vector space (Fig. 2.3). An additional coordinate is added to each of the three vectors and the angles between them are consequently increased.

More implicit definitions of concepts can be found, for example, in the field of word-sense disambiguation, especially in automatic identification of word senses (the so-called word-sense induction). For example, the approach to the word-sense induction known as *word clustering* involves grouping of semantically similar words. As such, they can reflect a specific meaning and can thus be considered as concepts. The determination of similarity between words can, for example, be based on the observation of the syntactic dependencies of particular words in actual texts (e.g. words acting as subjects of a verb, direct objects of a verb, adjectives of a noun etc.). In this manner, concepts can be implicitly defined as clusters of words that share a high amount of syntactic dependencies.

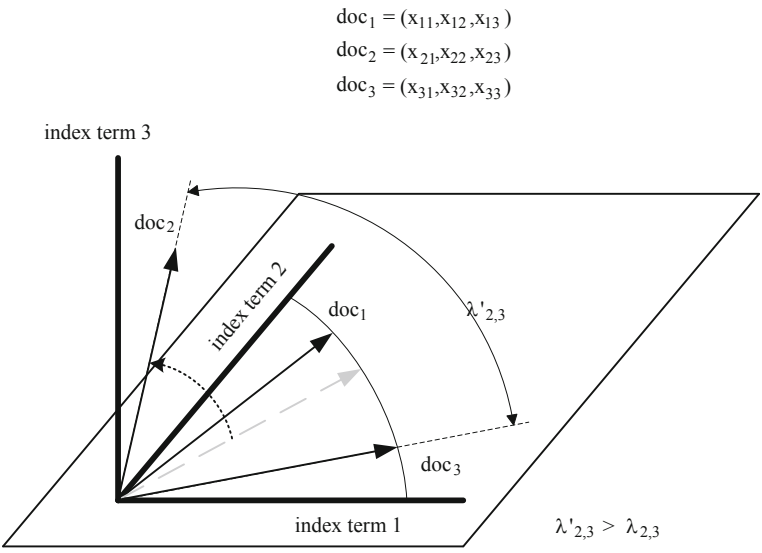

Fig. 2.3 Document representations after assigning a new index term to the Vector Space Model

Beside the one presented, many other approaches and algorithms exist in the field (see, for example, the survey of (Navigli 2009)), but what they all have in common is clustering—the grouping of words representing distinct meanings. These clusters can be interpreted as concepts. The exact definition of a concept, however, varies depending on the definition of similarity that a particular method uses for grouping words.

This subsection presented several possible definitions of concepts. The next section discusses the issue of how concepts can be organized and thus be made predictably available for use.

2.2 Organization

For concepts to be of any practical use for representing knowledge, they cannot be isolated; on the contrary, they must be linked with each other. In the beginning of this section, we introduce the relations among concepts as the key means of organizing concepts. As the linking of concepts eventually results in the formation of graph-like structures, the second part of this section presents the most important formalisms for representing such graphical organizations of concepts. The section concludes with examples of some specific concept organizations, for example when concepts are used as application data in databases and for document indexing and retrieval.

2.2.1 Relationships Among Concepts

The richness of the relationships linking concepts used in every day communication and the importance of identifying the underlying relationships between concepts are illustrated with the following example from Halladay and Milligan (2004).

> The statement "John has an IQ of 150" explicitly describes only a very simple relationship (i.e., that John has some attribute named IQ that equals 150). However, the statement assumes a myriad of other implicit relationships. These relationships include mundane things like IQ being the acronym for Intelligence Quotient, or that 150 is a value that precedes 151 and is preceded by 149, or that John is commonly a human male name, or that an IQ equal to 150 indicates a person of above-average intelligence, etc. However, without the context of all these relationships, the statement looses some of its fidelity or meaning. In fact, meaning is the sum-total of relationships.

The basic means for organizing concepts into knowledge structures are *semantic relations*. Semantic relations are meaningful associations between two or more concepts or entities (i.e. objects, instances or extensions of concepts) (Khoo and Na 2006).

The basic property of semantic relations is their *valence* (or "*arity*"), i.e. the number of concepts a semantic relation can associate. The valence of a relation is often expressed with the number of places, slots, fields or sides of the relation. Most often, the relations are binary, connecting two concepts. The relation "*give*" is, for example, a ternary relation, connecting the one who gives, the one who receives and that which is given. The relations with valence higher than two can be decomposed into binary relations. It was even suggested (Sowa 1984) that all relations can, in fact, be presented as concepts linked with a single (and the most primitive) "*link*" relation.

The quoted literature (e.g. Saussure 1916; Khoo and Na 2006; Stock 2010) distinguishes between two basic types of semantic relations: paradigmatic and syntagmatic relations. A *paradigmatic relation* is a relation between concepts that is independent of the actual use of the respective concepts (e.g. their occurrence in documents). For example, the concept "*guitar*" is inherently associated with the concept "*musical instrument*" by the paradigmatic relation "*is a kind of*" (the taxonomic relation). The hierarchical classification of paradigmatic relations is presented in Fig. 2.4. On the other hand, a *syntagmatic relation* is a relation between neighboring concepts in actual documents and as such, it only holds for an "ad-hoc" association of concepts in a particular document. An example of a syntagmatic relation is the relation "*play*" linking the concepts "*boy*" and "*guitar*" in the sentence "*That boy plays the guitar*".

Some of the formalisms for concept organization, such as object-oriented modeling languages and Semantic Web ontology languages, recognize a special type of relation, referred to as *the attribute*. Attributes denote the properties of concepts (or their extensions) and are usually represented as a feature of the entity

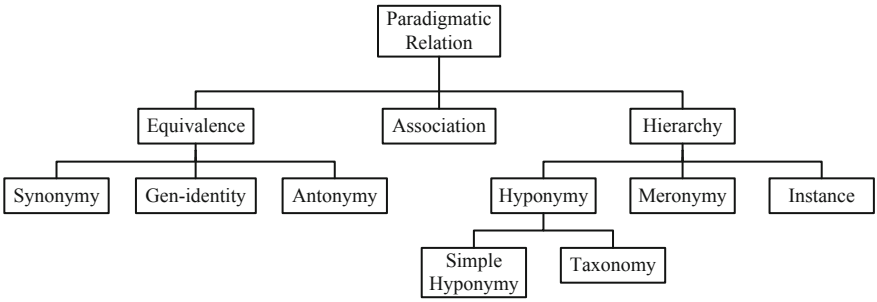

Fig. 2.4 The hierarchical classification of paradigmatic semantic relations (Stock 2010)

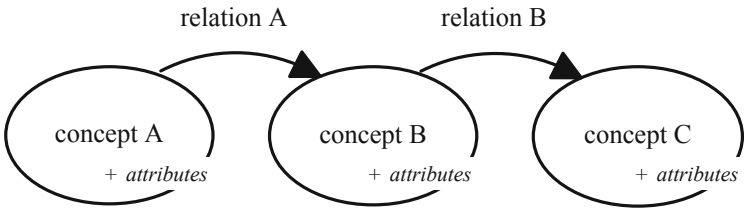

Fig. 2.5 The organization of concepts using relations and attributes

that is used to model concepts (or their extensions). Such representation of concepts using attributes in addition to semantic relations is illustrated in Fig. 2.5.

According to (Gomez-Perez and Corcho 2002), the attributes can be classified into the following four groups:

- *Class attributes* are assigned values that are attached to the concept and will therefore be the same for all instances of a concept.
- *Instance attributes* can be assigned different values for each instance (extension) of a concept.
- *Local attributes* are same-name attributes attached to different concepts.
- *Global attributes* can be applied to all concepts in a particular conceptual structure, for example, in an ontology.

The first two attribute types reflect the relationship between a concept and its instances. For example, the value of the attribute "chromosome number" appointed to the concept "human" is characteristic to all instances of this concept and can be thus considered a class attribute. On the other hand, the value of the attribute "age" is an instance attribute as it depends on a concrete person, for example, the next-door neighbor Susan, an instance of the concept "human".

The use of local and global attributes mainly depends on the representation requirements in actual applications. An example of a local attribute is the attribute "color". Although more diverse concepts within a particular conceptual structure can be related to this attribute, its value is local to a particular concept. An example

of a global attribute is the attribute "description", which holds a verbal explanation for every concept in a conceptual structure.

2.2.2 Graphical Organizations

In general, the use of semantic relations for linking concepts results in the organization of concepts in a graph-like structure. The three most common formalisms for describing such graphical structures are *conceptual graphs*, *concept maps* and *semantic networks*.

A conceptual graph (Sowa 1984) contains two kinds of nodes: concepts and conceptual relations. In the conceptual graph in the Fig. 2.6, concepts are presented with rectangles, and conceptual relations are presented with circles. Every edge in the conceptual graph links a conceptual relation to a concept. In the graph in Fig. 2.6, the concept *read*, thus, has an *agent* in the form of a person named *John*; a *theme*, the thing that is the object of the activity, in this case a *book*; and the *place*, where the situation is taking place, in this case the *living room*.

Concept maps are comprised of concepts and relationships between concepts. The relationships are expressed with words or phrases indicated on the edges linking concepts as shown in Fig. 2.7. The concepts and relationships in concept maps form propositions, or meaningful statements, "*about some object or event in the universe, either naturally occurring or constructed*" (Novak and Cañas 2008). Concepts in concept maps are organized hierarchically with the most general concepts arranged at the top of the map and more specific concepts placed bellow. When the propositions in a concept map are represented in a formal, computer-interpretable way, a concept map turns into a *semantic network* (Cañas and Novak 2009).

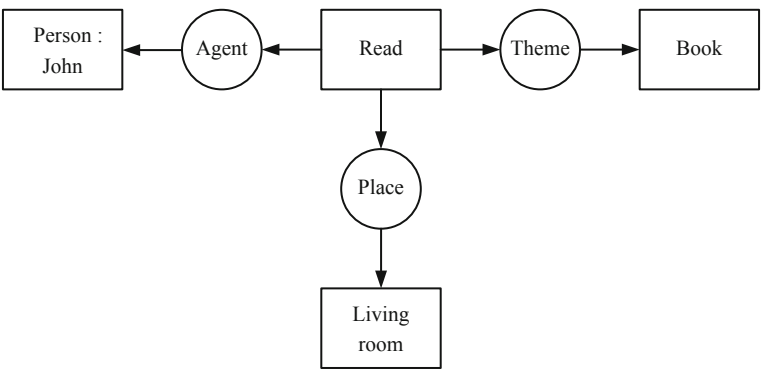

Fig. 2.6 Conceptual graph representing the proposition "*John is reading a book in the living room*"

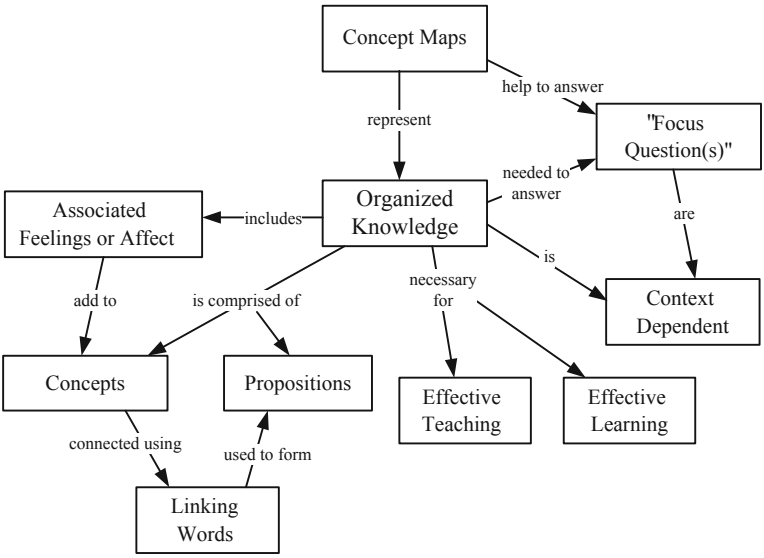

Fig. 2.7 A graphical organization of concepts in a concept map (Novak and Cañas 2008)

2.2.3 Specific Organizations of Concepts

2.2.3.1 Application Data in Databases

When intended to be used in actual applications, concepts are most often "stored" in computer databases and organized as data items together with respective cross-data relationships (Zellweger 2003). Some modeling techniques that follow the same approach are:

- entity-relationship model (Chen 1976),
- object-role modeling (Halpin 2006), and
- Unified Modeling Language (OMG 2012).

The above techniques share the same basic concept structure in which data items exhibit the associations among different neighboring data. The essence of grasping the meaning of the data in a database lies in the ability to assign an explanation to each of these associations, which provides a conceptual model for the application data (Zellweger 2003).

2.2.3.2 Indexing and Information Retrieval

Conceptual indexing offers one possible method of organizing concepts that are recognized in documents. The conceptual indexing method proposed by (Woods 1997) indexes phrases according to their conceptual structure instead of merely

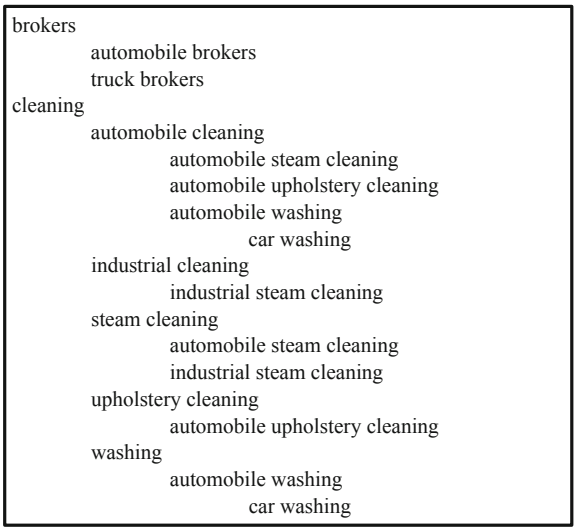

Fig. 2.8 A fragment of a conceptual taxonomy in a document related to automobiles (Woods 1997)

indexing them alphabetically. Each conceptual structure represents the way to compose the meaning of a phrase by combining several atomic elements, i.e. concepts.

An example of a conceptually indexed document is presented in Fig. 2.8. After parsing a phrase into one or more conceptual structures, the indexing system classifies the phrase according to the generality of its meaning. Such classification is achieved by utilizing the knowledge on the generality of relationships among individual elements of the phrase. For example, by utilizing the knowledge that a *car* is a sort of *automobile* and that *washing* is a kind of *cleaning,* the indexing system can determine that the phrase "*car washing*" represents a type of "*automobile cleaning*".

Another practical example of conceptual indexing is a thesaurus—the listing of words with similar, related, or opposite meanings. The "Joint INIS/ETDE thesaurus" (IAEA 2007), for example, contains structured information about the concepts in science and technology. Each record in the thesaurus consists of three components:

- a descriptor, i.e. a term identifying a concept;
- bibliographic data identifying the term entry date and corresponding remarks; and
- interrelationship indicators between individual concepts in the thesaurus. Three types of interrelationship indicators can be assigned:
  - preferential indicators (e.g. "used for"—UF),
  - hierarchical indicators (e.g. "broader term"—BT or "narrower term"—NT), and
  - an affinitive indicator ("related term"—RT).

PLUTONIUM
1996-01-24
UF dymac system
UF dynamic materials accountability system
BT1 actinides
BT1 transuranium elements
NT1 plutonium-alpha
NT1 plutonium-beta
NT1 plutonium-delta
NT1 plutonium-epsilon
NT1 plutonium-gamma
RT nuclear fuels
RT plutonium recycle

Fig. 2.9 An example from the Joint INIS/ETDE thesaurus

An example of an entry from the Joint INIS/ETDE thesaurus is presented in Fig. 2.9. The first two lines of the entry contain the term descriptor and the date of entry, while the remaining lines contain the relationships to other entries. The number following the interrelationship indicator (e.g. BT1) indicates the level (depth) of the relationship.

Conceptual indexing can also be conducted by ordinary users when annotating their documents for their efficient future retrieval. The conceptual indexing method proposed by Voss et al. (1999) is based on the manual marking textual elements in documents that are relevant to the user and could, therefore, also be relevant to others. The indexed concepts are not defined formally, therefore they must be interpreted in the context of their occurrences in the documents. The marked concepts can be organized by using two simple relations:

- the "*comprise*" relation, used for grouping several concepts into a new concept, and
- the "*associated*" relation, used when two concepts are considered to be closely associated, but not necessarily grouped into another concept.

The organization of concepts intended for describing the visual content of images is presented in (Jörgensen et al. 2001). The proposed conceptual structure organizes the visual attributes into four syntactic and six semantic levels. The attributes in the syntactic levels describe how an image is composed by using basic techniques and building blocks such as dots, lines, patterns and colors. The syntactic levels include the following four groups of attributes:

- *Type/technique* level contains the attributes that specify the type of image or the technique that was applied to create the image (e.g. color photograph).
- *Global distribution* level includes the attributes that classify the image based on its overall content that is determined by identifying the low-level perceptual characteristics such as color and texture (e.g. grey, blue, clear).

- *Local structure* level is involved in the categorization of the individual components extracted from the image (e.g. dots, lines, tones, texture).
- *Composition* level describes the specific layout of the basic elements in the image (e.g. symmetry, center of interest, leading lines and viewing angle), which are otherwise identified at the local structure level.

The attributes on the semantic levels describe the meaning of the elements in the image:

- *Generic Objects* and *Generic Scene* levels contain the attributes that describe the objects (e.g. tree, child, car) and scenes (e.g. city, landscape, portrait, indoor, outdoor), both of which can be recognized by using common knowledge.
- *Specific Object* and *Specific Scene* levels refer to the named entities (e.g. Albert Einstein, Eiffel Tower) and scenes (e.g. Paris, Times Square, Central Park). A more specialized knowledge is required for the recognition of specific objects and scenes compared to their generic counterparts.
- Attributes at the *Abstract Object* and *Abstract Scene* levels are used to describe what the individual objects in the image depict (e.g. angry woman) and what theme is represented in the image (e.g. sadness, happiness). The description at the abstract level is very demanding because it requires very specialized or even interpretative knowledge. Such description is therefore very subjective, as the choice of attributes might differ significantly for different describers.

The conceptual structure consisting of the introduced levels can be represented in the form of a pyramid, as shown in Fig. 2.10. The pyramidal shape illustrates

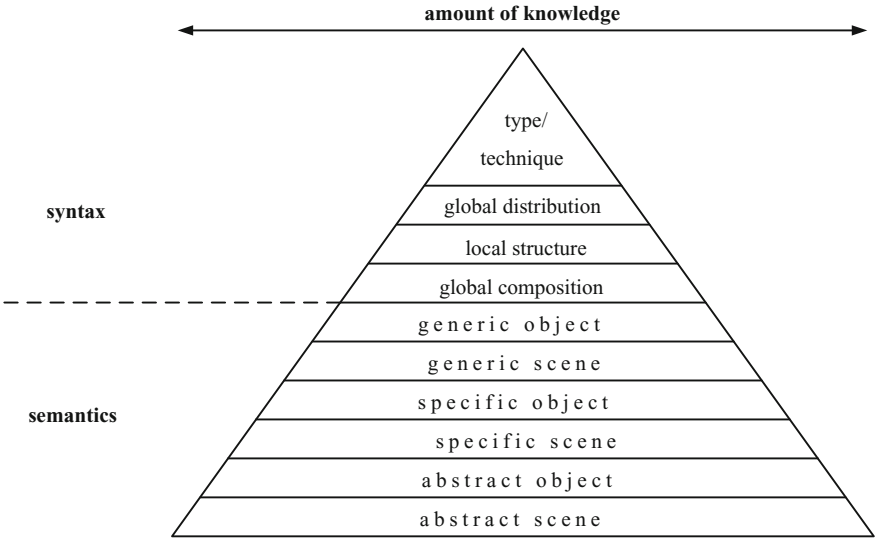

Fig. 2.10 Organizing concepts for describing the visual content in images based on the amount of knowledge needed for indexing (Jörgensen et al. 2001)

the volume of knowledge that is generally required for indexing images. The amount of knowledge required increases from the syntactic levels at the top to the semantic levels at the bottom of the pyramid. For example, to identify the individual objects in an image, more knowledge is required than to merely recognize the image type (e.g. color image). In addition, more knowledge is needed to identify a specific object or scene (e.g. face recognition, Central Park in New York City) than to recognize a generic object or scene (e.g. face detection, park).

The arrangement of attributes at various levels makes the model useful in many fields and for variety of indexing and retrieval methods. For example, the presented organization supports both automatic and manual indexing. Automatic indexing can be performed at syntactic levels where no specific world knowledge is required. On the other hand, the attributes at semantic levels used to categorize, describe and search for visual content can, in most cases, only be used by humans.

This section presented some general aspects of organizing concepts, including various forms of practical methods of organization applied to the fields of data storage, indexing and information retrieval. The following section is dedicated to the practical use of concepts organized in a uniform manner.

2.3 Concept Use

Various examples presented earlier in this chapter demonstrate that the same concepts can often be expressed by using different keywords. For example, if one says: "*I have a doctorate*" and "*I have a PhD*", these are two different statements, but they express the same concepts. In a typical searching scenario, the user is trying to retrieve the documents containing particular concepts of interest rather than exact keywords used in the query to refer to these concepts. As the concept-based search focuses on concepts rather than merely on words representing them, it presents a step ahead in the evolution of searching (Fig. 2.11). The concept-based search is, therefore, the main focus of this subsection related to concept use.

Three types of concept-based search systems are presented next: the Key-Concept is based on conceptual indexing, the Automated Generated Thesaurus

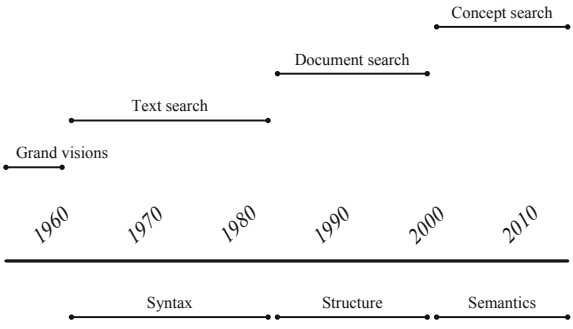

Fig. 2.11 Timeline of the searching evolution (Schatz 1997)

Approach provides alternate search terms in order to overcome the differences in terminology, and the Semantic Web search engines enable concept-based retrieval of the Semantic Web documents.

2.3.1 KeyConcept

The architecture of KeyConcept conceptual search engine is presented in Fig. 2.12 (Ravindran and Gauch 2004). In the indexing stage, the system applies conceptual indexing to the supplied documents. During the retrieval stage, the system ranks the indexed documents based on their similarity with the concepts (or the usual keywords) provided by the user.

The indexing process includes *classifier training* and *collection indexing*. In classifier training, the concepts in various sets of training documents are recognized and indexed by a traditional indexer using a modified *tf.idf* (*term frequency, inverse document frequency*) weighting method. The results of classifier training are vectors containing weighted terms representing distinct concepts that had been recognized during the training stage. The method for calculating weights is presented in Fig. 2.13. The weight of a term depends on (i) the frequency of the term's occurrence in the training documents for the particular concept the term represents, (ii) the rarity (or inversed frequency) of the term in the training documents for all concepts, and (iii) the frequency of the individual documents in the training set containing the term representing the particular concept.

During *collection indexing*, the conceptual indexer processes new documents using a Vector Space Model (Salton and Wong 1975) (see Sect. 2.1.2).

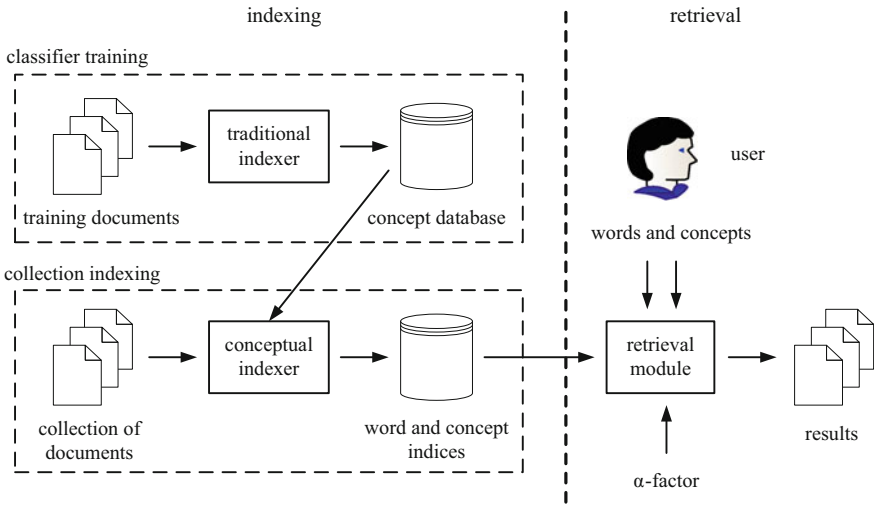

Fig. 2.12 The architecture of KeyConcept (Ravindran and Gauch 2004)

$$wt_{i,j} = tf_{i,j} * icf_i * cdf_{i,j}$$
$$icf_i = \log(n/cf_i)$$
$$cdf_{i,j} = \log(cdf_j/df_i)$$

$wt_{i,j}$ (term weight) = the weight of the term *i* representing the concept *j*
$tf_{i,j}$ (term frequency) = the frequency of term *i* in the training documents for concept *j*
icf_i (inverse concept frequency) = the rarity of the term *i* in the training sets for all concepts
$cdf_{i,j}$ (concept document frequency) = the frequency of the individual training documents in the training set containing the term *i* as a representative of the concept *j*

n = total number of concepts
cf_i = number of concepts containing term *i*
cdf_j = number of training documents for concept *j*
df_i = number of training documents containing term *i*

Fig. 2.13 The modified *tf.idf* weighting method used in KeyConcept

The supplied documents are classified by comparing their representative vectors to concept vectors that were computed in classifier training. The results of the collection indexing are indices that represent the similarity between a particular document in the collection and the concept vectors. The concept indices for each document are stored in the "word and concept index" (WCI) database which, as its name suggests, also contains standard word indices.

The retrieval is carried out by searching the WCI database for keywords or concepts that were provided by the user. The ranking of search results is computed by considering the relative importance of concept matching in relation to word matching, which is provided by the configurable α-factor.

2.3.2 Automated Generated Thesaurus Approach

Thesauri can be an efficient tool for concept-based retrieval in the text domain. The Automated Generated Thesaurus Approach (AGTA) (Chen et al. 1998) provides an ability to fine tune the keywords a user had (or should have had) in mind when forming a particular query. AGTA carries this out in several stages:

- *Document collection* includes collecting a set of documents in a subject domain serving as the thesaurus base.
- *Automatic indexing* involves the identification of terms (i.e. "*subject descriptors*") appearing in the document collection using the automatic indexing technique proposed by Salton (1989).
- In *co-occurrence analysis*, first term frequency (*tf*) and inverse document frequency (*idf*) are calculated (see Sect. 2.3.1). The two values are used to assign weights to each term in a document in order to represent the term's level of importance.

Cluster analysis then creates a network of terms with weighted connections, describing the similarity among terms.

- In the *associative retrieval* stage, the Hopfield algorithm (Hopfield 1982) is used to retrieve the terms that are related to the input term provided by the user. The algorithm first "activates" the neighbors of the input term, which are most strongly associated with the term itself. The algorithm then combines their weights and repeats the activation process on their neighbors. As the algorithm is iterative, this activation process "spreads" away from the input term and eventually fades away. The activation weights of the terms farther away from the input term are, namely, gradually decreasing and thus these terms are eventually excluded from the activation process.

Due to the fact that associative retrieval provides synonym terms or terms that have very similar meaning, these terms can be considered as representations of the same concept. The retrieved terms can thus be used to overcome vocabulary differences, for example in scientific information retrieval or in order to alleviate the problem of information overload when searching through huge information sources, such as the Web. For example, by using the terms suggested by the thesaurus, a user can enhance keyword-based web search to retrieve more relevant results.

2.3.3 Semantic Web

Before addressing the concept-based search on the Semantic Web, we shall briefly describe the latter. The Semantic Web (Berners-Lee et al. 2001) is the vision of Web of the future with the structure of information that is understandable to computers, so the latter can perform many tasks instead of humans, for example finding, sharing, and combining information. The technologies that constitute the Semantic Web are organized in the so-called Semantic Web stack. The stack with outlined key technologies is shown in Fig. 2.14.

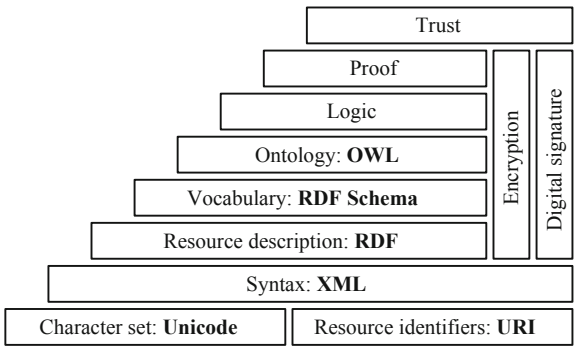

Fig. 2.14 The Semantic Web technology stack

The essence of the Semantic Web is *Resource Description Framework* (RDF) (RWG 2012a), a language for describing resources in the form of triplets. The triplets consist of a subject, an object and a predicate associating the former two. For example, the predicate "title" associates this book (the subject) with its title "Concepts, Ontologies, and Knowledge Representation" (the object).

The layers below the resource description layer provide the means to encode RDF triplets without imposing any semantic constraints on the selection of the subject, predicate and object.

- *eXtensible Markup Language* (XML) (W3C 2012), a language designed to transport and store data, provides the basic syntax for encoding the triplets. RDF documents recorded in the syntax of XML can easily be exchanged between computers and applications.
- The content in XML is encoded in Unicode standard that provides a character set for the representation of text in the majority of the presently used writing systems.
- The resources described by the RDF triplets can be any objects expressed with a *Uniform Resource Identifier* (URI) (Berners-Lee et al. 2005). URI is a string of characters that identifies a resource in a network.

The layers above the resource description layer comprise technologies that provide meaning to the RDF statements:

- RDF Schema (RWG 2012b) uses RDF language to define the basic application-specific vocabulary that is used for describing resources with the means of classes, subclasses and properties. The classes are often viewed as concepts, as they combine a set of individual resources (objects) with common properties.
- *Web Ontology Language* (OWL) (OWL 2009) extends the vocabulary of RDF Schema with the constructs that enable the description of more advanced relationships among the classes, thus enabling the construction of ontologies.

The upper three layers in the Semantic Web stack are not yet fully standardized and implemented. As such, they currently mostly represent the ideas that are supposed to be realized in order to entirely implement the Semantic Web. These layers include:

- the Logic layer containing the means to make inferences based on knowledge represented in ontologies;
- the Proof layer that executes the rules defined in the Logic layer; this layer enables the drawing of conclusions from given sets of facts and thus acquiring new knowledge from the knowledge already adopted; and
- the Trust layer containing the decision making mechanisms to differentiate whether to trust the given proof from the bottom layers. In addition, the mechanisms of cryptography, such as digital signature and encryption, may also be used to ensure privacy and verify that the Semantic Web statements originate from a trusted source.

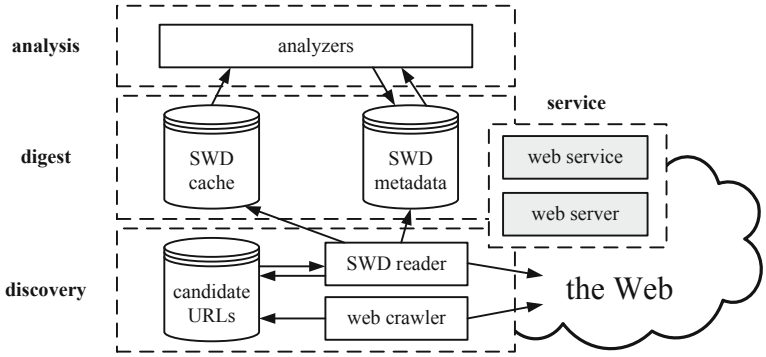

Fig. 2.15 The architecture of the Semantic Web search engine Swoogle (Finin et al. 2005)

Semantic Web technologies enable searching that is not limited merely to keyword matching, but can in fact be based on concepts. Examples of such Semantic Web search engines include GoPubMed (Doms and Schroeder 2005; GoPubMed 2012), *Semantic Web Search Engine* (SWSE) (Hogan et al. 2011; SWSE 2012; hakia 2012; Watson 2012; Sindice 2012) and Swoogle (Ding et al. 2004; Finin et al. 2005).

The architecture of Swoogle, for example, comprises several components (Fig. 2.15):

- The *discovery component* consists of *crawlers* that discover candidate *Uniform Resource Locators* (URLs) referencing *Semantic Web Documents* (SWDs). SWDs are all documents containing the RDF triplets.
- The *digest component* caches SWDs from the Web. The component then creates their corresponding metadata as well as the metadata describing individual concepts (also referred to as SWTs—Semantic Web Terms) contained in the SWDs. In addition, this component also identifies the relations among individual SWDs and SWTs.
- The main task of the *analysis component* is to classify and rank the cached SWDs and SWTs by their importance. The ranking algorithm used by Swoogle is a modified version of Google's PageRank algorithm (Page et al. 1999), adapted to account for the semantics of links in SWDs.

Searching for SWDs and their relationships with other SWDs and SWTs is possible through the web interface (Swoogle 2012) as well as through web services.

Although concepts and ontologies as such are presented in separate layers in the Semantic Web technology stack, in practice there is often no clear distinction where the use of uniformly organized concepts stops and the use of ontologies begins. The criteria for classifying the organization of concepts in an ontology are discussed in the next chapter.

References

Baader F, Nutt W (2002) Basic description logics. In: Baader F et al (eds) The description logic handbook. Cambridge University Press, Cambridge, pp 47–100

Berners-Lee T, Hendler J, Lassila O (2001) The semantic web. Sci Am 284(5):34–43

Berners-Lee T, Fielding R, Masinter L (2005) Uniform resource identifier (URI): generic syntax, RFC 3986. http://tools.ietf.org/html/rfc3986. Accessed 25 Sep 2012

Boyer CB (1959) The history of the calculus and its conceptual development. Courier Dover Publications, New York

Cañas AJ, Novak JD (2009) What is a concept map? http://cmap.ihmc.us/docs/conceptmap.html. Last update 28 Sep 2009. Accessed 25 Sep 2012

Chen H, Houston AL, Sewell RR, Schatz BR (1998) Internet browsing and searching: user evaluation of category map and concept space techniques. J Am Soc Inform Sci 49(7):582–603

Chen PP (1976) The entity-relationship model—toward a unified view of data. ACM Trans Database Syst 1(1):9–36

Ding L, Finin T, Joshi A, Pan R, Cost RS, Peng Y, Reddivari P, Doshi VC, Sachs J (2004) Swoogle: a search and metadata engine for the semantic web. In: Proceedings of the thirteenth ACM international conference on information and knowledge management, Washington, DC, Nov 2004

Doms A, Schroeder M (2005) GoPubMed: exploring PubMed with the gene ontology. Nucleic Acids Res 33:783–786

Fellbaum C (ed) (1998) WordNet: an electronic lexical database. MIT Press, Cambridge

Finin T, Ding L, Pan R, Joshi A, Kolari P, Java A, Peng Y (2005) Swoogle: searching for knowledge on the semantic web. In: Proceedings of the national conference on artificial intelligence (AAAI), Pittsburgh, 2005

Frege G (1892) Über Sinn und Bedeutung. In: *Zeitschrift für Philosophie und philosophische Kritik*. English edition: Frege G (1980) On Sense and Reference (trans: Black M). In: Geach P, Black M (eds) *Translations from the Philosophical Writings of Gottlob Frege*, 3rd edn. Blackwell, Oxford

Gomez-Perez A, Corcho O (2002) Ontology languages for the semantic web. IEEE Intell Syst 17(1):54–60

GoPubMed® (2012) http://www.gopubmed.org. Accessed 25 Sep 2012

hakia.com (2012) http://www.hakia.com/. Accessed 25 Sep 2012

Halladay S, Milligan C (2004) The application of network science principles to knowledge simulation. In: Proceedings of the 37th annual Hawaii international conference on system sciences, Big Island, 5–8 Jan 2004

Halpin T (2006) Object-role modeling (ORM/NIAM). In: Bernus P, Mertins K, Schmidt G (eds) Handbook on architectures of information systems. International handbooks on information systems. Springer, Heidelberg, pp 81–103

Hogan A, Harth A, Umbrich J, Kinsella S, Polleres A, Decker S (2011) Searching and browsing linked data with SWSE: the semantic web search engine. J Web Semant 9(4):365–401

Hopfield J (1982) Neural networks and physical systems with emergent collective computational abilities. Proc Natl Acad Sci USA 79(8):2554–2558

International Atomic Energy Agency (2007) Joint thesaurus, Part I + II. ETDE/INIS joint reference series no. 1 (Rev. 2). http://www-pub.iaea.org/MTCD/publications/PDF/JRS1r2_web.pdf. Accessed 25 Sep 2012

Jörgensen C, James A, Benitez AB, Chang SF (2001) A conceptual framework and empirical research for classifying visual descriptors. J Am Soc Inf Sci Technol 52(11):938–947

Kant I (1800) Logik. English edition: Kant I (1988) Logic (trans: Hartman RS, Schwarz W). Dover Publications, Mineola

Khoo C, Na J-C (2006) Semantic relations in information science. Annu Rev Inform Sci Technol 40(1):157–228

Locke J (1690) An essay concerning human understanding. Oxford University Press, New York, (1975)
Miller GA (1995) WordNet: a lexical database for English. Commun ACM 38(11):39–41
Navigli R (2009) Word sense disambiguation: a survey. ACM Comput Surv 41(2). doi: 10.1145/1459352.1459355
Novak J, Cañas A (2008) The theory of underlying concept maps and how to construct them. Technical report IHMC CmapTools 2006-01 Rev 01-2008, Florida Institute for Human and Machine Cognition
Ogden CK, Richards IA (1923) The meaning of meaning. Harcourt, Brace & Co, New York
Object Management Group (2012) UML® resource page. http://www.uml.org/. Accessed 25 Sep 2012
OWL Working Group (2009) OWL 2 web ontology language document overview. W3C recommendation 27 October 2009. http://www.w3.org/TR/owl-overview/. Accessed 25 Sep 2012
Page L, Brin S, Motwani R, Winograd T (1999) The PageRank citation ranking: bringing order to the web. Technical report, Stanford InfoLab
Petersen W (2007) Representation of concepts as frames. In: Skilters J, Toccafondi F, Stemberger G (eds) Complex cognition and qualitative science. The baltic international yearbook of cognition, logic and communication, vol 2. University of Latvia Press, pp 151–170
Princeton University (2010) About WordNet. http://wordnet.princeton.edu. Accessed 25 Sep 2012
Ravindran D, Gauch S (2004) Exploiting hierarchical relationships in conceptual search. In: Proceedings of the thirteenth ACM international conference on information and knowledge management, Washington, DC, 2004
RDF Working Group (2012) RDF—semantic web standards. http://www.w3.org/RDF/. Accessed 25 Sep 2012
RDF Working Group (2012) RDF vocabulary description language 1.0: RDF Schema (RDFS). http://www.w3.org/2001/sw/wiki/RDFS. Accessed 25 Sep 2012
Salton G (1989) Automatic text processing: the transformation, analysis, and retrieval of information by computer. Addison-Wesley, Reading
Salton G, Wong A (1975) A Vector Space model for automatic indexing. Commun ACM 18(11):613–620
Saussure F de (1916) Cours de linguistique générale. English edition: Saussure F de (1983) Course in General Linguistics (trans: Harris R). Open Court, La Salle
Schatz B (1997) Information retrieval in digital libraries: bringing search to the net. Science 275(5298):327–334
Schopenhauer A (1851) Parerga and paralipomena. English edition: Schopenhauer A (1974) Parerga and paralipomena: short philosophical essays (trans: Payne EFJ). Oxford University Press, New York
Sindice—the semantic web index (2012) http://www.sindice.com/. Accessed 25 Sep 2012
Smith B (2004) Beyond concepts: ontology as reality representation. In: Varzi AC, Vieu L (eds) Proceedings of the international conference on formal ontology and information systems, Turin, 4–6 Nov, pp 73–84
Sowa JF (1984) Conceptual structures: information processing in mind and machine. Addison-Wesley Publishing, Reading
Sowa JF (2000) Ontology, metadata, and semiotics. In: Conceptual structures: logical, linguistic, and computational issues. Lecture notes in computer science, vol 1867. Springer, Berlin, pp 55–81
Stock WG (2010) Concepts and semantic relations in information science. J Am Soc Inf Sci Tec 61(10):1951–1969
Swoogle Semantic Web Search Engine (2012) http://swoogle.umbc.edu/. Accessed 25 Sep 2012
SWSE—Semantic Web Search Engine (2012) http://swse.deri.org/. Accessed 25 Sep 2012
Voss A, Nakata K, Juhnke M (1999) Concept indexing. In: Proceedings of the international ACM SIGGROUP conference on Supporting group work, Phoenix, 14–17 Nov 1999, pp 1–10
W3C (2012) Extensible Markup Language (XML). http://www.w3.org/XML/. Accessed 25 Sep 2012

Watson Semantic Web Search (2012) http://kmi-web05.open.ac.uk/WatsonWUI/. Accessed 25 Sep 2012

Woods W (1997) Conceptual indexing: a better way to organize knowledge. Technical report, Sun Microsystems, Mountain View

Zellweger P (2003) A knowledge-based model to database retrieval. In: Proceedings of the international conference on integration of knowledge intensive multi-agent systems, Boston, 30 Sep–4 Oct 2003, pp 747–753