The Open University

KMI

Annotating Knowledge Resources to Support Learning

First Year Probation Report

2009 **Petr Knoth**

Contents

Abstract

This work investigates and reviews state-of-the-art approaches to automatic organization of unstructured digital content. Given the review of the current challenges, the research questions address a vital issue of automatic metadata generation with the focus on link generation and link typing based on the analysis of content. The motivation and the potential impact these methods may generate both in the general context and in the context of digital learning repositories are discussed. Finally, a detailed research plan and the current results are presented.

Chapter 1

Introduction

The amount of new information accessible online is increasing rapidly. This illustrates the need to organize and maintain the information efficiently. However, it will be demonstrated that keeping the growing amounts of information well-organized according to a certain criterion may not be feasible for humans and thus the development of automatic methods is inevitable. In this report, the state-of-the-art in methods for organization of unstructured textual content is reviewed. Based on the gap analysis of the state-of-the-art, research questions and a research proposal are formulated. The proposal aims at development of novel link generation and link typing methods that are capable of, but not limited to, improving the organization and navigation over independent information resources, such as articles in digital repositories, by using entirely automatic methods.

1.1 Motivation and Research Problem

1.1.1 A motivating scenario

Let us now introduce the research problem on a motivating example. We will assume that in 1995 we have written an article about global warming where we described the current evidence supporting the view that global warming is highly influenced by the amount CO_2 in the atmosphere. The article analyzed the contemporary political views on how to reduce the exhalations in different countries and in its conclusions was referring to a global warming prognosis for the next 50 years performed by scientists using state-of-the-art methods. When the article was written, it was reviewed by an independent referee and later published on the Internet by a trustworthy source. Let us also assume that the publisher had been wise and had asked us before the article was published to provide certain metadata about our article. More specifically, the publisher required us to submit keywords, to classify the article to a taxonomy and to provide information about related sources of information, for example sources that support our hypothesis and sources where the learner may find more information. The metadata made it easier for people to retrieve the article and to navigate to related resources.

Since the article was published, many new papers discussing similar issues have appeared and lots of research have been carried out. However, the new

information growth caused some troubles. Since the domain and its terminology have evolved, it may be now in few cases possible to find more descriptive keywords, such as a specific class of the method used for the prediction, than those identified at the time of writing. This causes that the article can be no longer retrieved as easily at appropriate times as it used to be. As the amount of content had also increased, the publisher had to revise the taxonomy for the articles in the global warming domain. Due to the fact that many articles had to be inspected and reclassified by a domain expert, this procedure was very difficult and time consuming. The biggest issue arose with the metadata that provided relations to other useful sources. While some of the sources indicated by us at the time of writing are no more available, other sources which have emerged since the time of writing are not known to the publisher. As a result, this information is lost.

The motivating scenario indicates that keeping information up to date may be a tedious task. The increased need for coordination has been also observed in open collaborative knowledge-management systems, such as Wikis [Kittur et al., 2007] which typically use a slightly different maintenance policy than the one described in the mentioned scenario. In particular, it has been noted that the effort necessary for the maintenance of the information is not directly proportional to the amount of information stored, but rises faster than direct proportionally to the amount of information being added.

So, how can we do better than in the motivating scenario? First of all, it is important to address the question of what kind of support is needed. Surely, if the keywords, the classification and the generation of links were derived automatically, the maintenance of such a system would become almost effortless. While this would make a significant difference to the publisher, there is less benefit for the user of such a system, who is typically embodied by a learner investigating a certain area of a given domain. Therefore, we should now focus on how the learner and their learning processes can be made more efficient.

Since 90's there seems to be a gradual shift from traditional hard copy text approaches to more complex blended learning approaches where learners can take control over their learning strategy by using links that are interconnecting content. For example, on Wikis learners typically use links that connect a textual entity, mostly a terminological unit, to a page which describes it. However, these collaborative efforts are based on a *centralistic* view allowing the existence of only one instance of an article describing a certain concept. Therefore, this approach does not fit well when one wants to publish their own view on a given problem, their own research results, their own methodology for solving a problem etc. As a matter of fact, most of the information publishing tasks being carried out in our World are in fact entirely *distributed*. For example, many books for teaching history which differ in a variety of aspects have been published. Even books teaching sciences, such as mathematics or physics, and focusing on the same problem domain may differ in their narrative structure, focus, level of detail, intended target audience, language and other factors.

In this light, we see that current systems should help learners in discovering and maintaining links from the resource or passage currently being viewed to resources or passages that are related. An important aspect of such linking is that both the system and the learner should be aware of the type of the link. The idea of link typing is not new. In 1983 Trigg developed a link taxonomy [Trigg, 1983] which classifies links according to their semantics. However, computer

systems are currently unable to automatically derive most of these link types. Some of the important link types which may be possible to derive automatically were identified later in [Allan, 1996] by James Allan who mentions, for example, *revision* links, *summary* and *expansion* links, *equivalence* links, *comparison* and *contrast* links, and *tangent* and *aggregate* links. There are other important types of links not mentioned in this study, but highly important in the educational context, such as *citation* links and *prerequisite* links.

Let us now conclude by explaining how we would envisage our scenario described at the beginning of this section to work if all the metadata including the typed links among resources or passages were updated always just in time. The theoretical constraints of this approach will be discussed in section 1.1.2.

Since our article was published in 1995, the metadata have been regularly updated regardless of the cost. Furthermore, links to related resources or passages of related resources were annotated by their type. Because of this, it is easy to retrieve the article using search engines and domain-specific web directories. When a learner accesses and reads our global warming resource, he/she is offered links to related resources or even passages in related resources. These links do not necessarily commence from the whole resource, but may, for example, link a paragraph in the resource to a paragraph in another resource. The possibilities of related content links presented to the learner are demonstrated on the following examples: 1) three other authors also published papers about the same topic, but from a much broader perspective. The system displays them and indicates the type of the relationship. 2) The system also shows that there are a few more in depth studies on particular aspects discussed in our article, such as the political issues and the mathematics behind the prediction model. 3) In addition, the system provides a *prerequisite* link to a lecture in statistics that may serve the learner in understanding the mathematics behind the prediction model. 4) The system highlights that there is also a *tangential* link to a paper published five years after our article that is criticizing the impreciseness of the mathematical model and is questioning the validity of such long term predictions. 5) Finally, another link relates our lecture to a blog submission where our paper was summarized.

1.1.2 Theoretical constraints

In this section, the time constraints for annotation of resources will be investigated. Metadata for the description of resources will be divided into three classes according to their type. Maximum time complexities for their provision will be then given.

In the context of systems providing materials that can be used for learning, we can also refer to a resource as a *learning object (LO)*. The Learning Technology Standards Committee of the IEEE Computer Society defines a learning object as any entity, digital or non-digital, that may be used for learning, education, or training [LOM, 2005]. As most of the issues discussed in this report are valid for both learning objects and resources, the term *resource* will be rather used. By resource, we will refer to any textual resource in a digital format. Resources can be used for learning, but are not limited to this use. They may epitomize, for example newspaper articles, blog submissions or scholarly articles. We will refer to repositories of resources as digital repositories.

To organize resources, it is necessary to annotate them with appropriate

metadata. An essential requirement for the metadata is that they are described in a machine readable way. Another requirement is that the metadata description follows a particular standard, thus allow interoperability (section 2.1). We divided metadata into three distinct classes according to their type. One of the well-known metadata scheme IEEE Learning Object Metadata (LOM) [LOM, 2005] was used to validate that each metadata field fits into a class and that each class is represented by a metadata field. The following metadata types were identified:

1) Metadata fields describing the content of a resource. This type of metadata specifies concepts, such as the name of an author, title of a resource or a set of keywords used.

2) Metadata fields classifying a resource using a taxonomy. This type is used to associate a resource with a coarse grained structure of the subject domain.

3) Metadata fields connecting two resources usually by a semantic relation.

Manual provision of type 1 metadata requires an annotator to have only the knowledge about a given resource. Metadata of type 2 require to understand the subject domain (i.e. to understand the resource context). Providing type 3 metadata requires in the extreme case the understanding of all resources available in a digital repository and the checking whether a semantic relation holds.

We will now explore the maximum time complexity for the provision of the previously mentioned metadata types. Let t_1 denote the maximum time needed to access, view and broadly understand a resource. Let h denote the number of nodes/topics in a classification taxonomy and let t_2 denote the maximum time needed to check whether a given resource should be associated with a node in the taxonomy. Finally, let n be the number of resources available in a collection. Then, the maximum time t_{max} needed to provide type 1 metadata is:

$$t_{max} = t_1.n \Rightarrow t(n) = O(n), \tag{1.1}$$

thus the time complexity is linear with respect to the number of resources available. The maximum time needed to generate type 2 metadata is:

$$t_{max} = (t_1 + t_2.h).n \ \Rightarrow t(n) = O(n) \tag{1.2}$$

The maximum time is given by the time of accessing/understanding a resource plus the time of associating the resource to a taxonomy times the number of resources available in the collection. The complexity is still linear with respect to the number of resources available, but the actual time required for annotation rises also linearly with respect to the number of nodes in the taxonomy. It is important to note that we assume that a resource can be associated to any number of the taxonomy nodes. If exactly one taxonomy node from a tree shape taxonomy is selected, the maximum cost of association would be logarithmic with respect to the number of taxonomy nodes. The base of the logarithm is given by the branching factor of the tree structure.

Finally, the maximum time spent in deriving type 3 metadata is given by the following expression:

$$t_{max} = (t_1.n).[t_1.(n-1)] \Rightarrow t(n) = O(n^2) \quad (1.3)$$

The equation states that for the generation of links specifying one type of a binary semantic relation it is necessary to access all resources and to take into account all remaining resources for each of them. The time complexity is quadratic with respect to the number of resources stored in the repository. For simplicity, the equation assumes that the annotator does not have any memory and thus needs to access, view and understand a particular resource each time again. While it is possible, in practise, that an annotator (with a memory) can work much faster than predicted by the equation, it is not possible to avoid the quadratic number of comparisons with respect to the number resources.

1.1.3 Discussion

Based on the equations, we now discuss the feasibility of manual annotation. It can be seen that when t_1 is small, providing type 1 metadata may be for human annotators relatively effortless. Generating type 2 metadata may be still possible to perform in case t_2 and h are small or if the task is formulated as a *one-of* problem (exactly one taxonomy node is selected). However, specifying type 3 metadata can be performed by humans only for a very limited amount of resources. For example, if we assume that accessing and understanding a resource takes one minute, the annotation of a collection of 100 resources can take up to 165 hours. Furthermore, binary linking of resources is the most difficult metadata type to maintain as adding a new resource to a collection would have typically much higher frequency than changing a classification taxonomy or a set of possible keywords. Last, in certain type of collections, such as multilingual collections, it may be for humans extremely difficult to perform such task.

To conclude, human performed link generation does not scale up and can become infeasible even for very small collections. This makes link generation also theoretically unsuitable for collaborative approaches[1] which can be well applied to type 1 and type 2 metadata. A predominant approach is to generate links based on text analysis of documents or their type 1 or type 2 metadata. Current computer systems are capable of generating semantic similarity links in repositories containing up to one million of resources [Manning et al., 2008] when all possible pairs are checked, thus by algorithms with $O(n^2)$ complexity. For larger repositories, approximations calculated by algorithms with lower complexity can be used (see [Manning et al., 2008]). As humans are unable to carry out the link generation task, even algorithms with a relatively low precision and recall, in comparison to type 1 and type 2 metadata generation methods, are of real value.

Many of the problems of automatic metadata extraction have been already addressed by a variety of approaches with different levels of success. Using approaches that analyse human language to extract metadata, task 1) has been in educational settings addressed, for example, by information extraction in [Bateman et al., 2007]. Task 2) can be seen as a text classification task [Sebastiani, 2002]. Finally, task 3) represents a problem of automatic link generation [Wilkinson and Smeaton, 1999]. While there is a great need for link generation and in particular for link typing methods, which would provide people with

[1] This may be possible only for resources that have a high visit frequency by domain experts

access to information as in the motivation scenario, the research field is in comparison to keyword extraction and text classification approaches still relatively unexplored. An exception may be generation of links based on the traditional semantic similarity measures. The state-of-the-art in link generation will be presented in section 2.2.4.

1.2 Outline

So far, the problem domain has been described and it was argued why metadata extraction and in particular link generation is an important research problem. The state-of-the-art in the field will be reviewed in chapter 2. The research questions, the expected contribution and the proposed approach are presented in chapter 3. Finally, in chapter 4, the work carried out so far is demonstrated.

Chapter 2

Literature Review

This chapter reviews the state-of-the-art in the field related to the problematics described in the previous chapter. In section 2.1, we discuss what metadata are currently being used by metadata standards to allow sharing of content. Section 2.2 reviews automatic approaches to content organisation, in particular approaches to: keywords extraction (section 2.2.1), text classification (section 2.2.2), learning pathways (section 2.2.3), and link generation and typing (section 2.2.4). Section 2.3 provides the gap analysis and shows the direction for the research proposal presented in the next chapter.

2.1 Interoperable resources

The ability to communicate and share resources is an essential requirement for digital repositories, as discussed in section 1.1.2. As a result of this, a number of metadata standards for the description of resources arose. Perhaps the most well-known metadata standard is Dublin Core [Pidgin and Baker, 1997], which defines a set of elements, such as title, language, subject, coverage or relation, for the description of resources. It is interesting to note that despite the Dublin Core's simplicity, it already contains all three types of metadata identified in section 1.1.2. For example, the *subject* field should contain keywords, the *coverage* field should ideally contain a value from a controlled vocabulary and the *relation* field should contain a reference to a related resource. The set of elements defined by Dublin Core may not be sufficient in all application domains and under all circumstances. One of the most important standards in the educational domain is Learning Object Metadata (LOM) [LOM, 2005]. This standard, current issues and the main tradeoffs in the design of a metadata standard for digital resources will be now discussed.

The main role of metadata standards is to catalogue important information about resources in a unified way. Apart from traditional retrieval facilities, such as keyword search, or hierarchical classification, the learning community tries to find solutions to relatively challenging problems. For example, Brooks explains that the ability for a software system to dynamically, and ideally in real-time, assemble a course from learning resources is an important goal within the educational technology research community x[Brooks and Mccalla, 2006]. His interest is not only in assembling courses from resources stored in a central database,

but he envisions automatic course assembly from distributed repositories.

Brooks argues that the current e-learning standards specifications, considering LOM in the first instance, are both too restrictive in the variety of metadata they capture and too lax in how they express the structure of such metadata. In addition, as mentioned in [Simon et al., 2005], many of the current learning systems support only a few fields described by LOM and most of them do not even support any external format. This makes the vision of computer agents retrieving and collecting learning resources from distributed repositories harder to achieve. There have been quite a few EU funded projects, such as the Metadata for Architectural Contents in Europe (MACE) [Cress et al., 2009], which have addressed the issue of interoperability by harmonizing distributed digital repositories with respect to the LOM standard. The harmonization process usually requires quite a lot of effort. One of the reasons is that the LOM standard does not precisely specify the types of the various fields and their interpretation.

There have been also other criticisms of the LOM standard. For example, Brooks criticizes the centralization of the metadata and claims that for the retrieval of learning resources are not only essential summarized characteristics of resources, but also situational metadata, defining user's background, outcome information and user's interaction with a learning object. So far, many researchers investigated the use of personalized user profiles [Keenoy et al., 2004; Brusilovsky et al., 2007] and applied them to digital repositories. However, we can argue whether criticizing the LOM standard for such incompleteness is in place, because the nature of dynamically collected metadata about a user is conceptually very different from what LOM standard wants to achieve. Perhaps, the best option is to keep the LOM standard as it is and develop the specification of a personalization model in terms of a different standard.

It is possible to see that the LOM standard, and other metadata standards, face to a tradeoff between being too *specific* or being too *general*. The *specificity* causes the standard to define very clear rules for developers that are necessary for ensuring interoperability. Such rules may specify the semantics or even intended interpretation of the various metadata fields, however this is not the case of LOM. At the same time, it is important to keep the standard *general* enough to allow extensions. According to the current specification, the LOM standard can be combined with other forms of knowledge representation. For example, metadata records can be linked to external shared conceptualizations [Urbán and Barriocanal, 2003]. The justification for allowing this linking is following. The LOM standard explicitly states that it does not specify how a learning technology system represents or uses a metadata instance for a learning object. More specifically, it is concerned only with the syntax form of the data it represents and it does not provide any formal semantics. To conclude, in order to allow interoperability between distributed repositories any standard should be as specific as possible. On the contrary, in order to allow the use of the standard across a variety of unanticipated applications, the standard should be general enough.

To give an example, the authors of [Urbán and Barriocanal, 2003] recommend to use the LOM's classification element to declare links to ontology terms. This, together with the purpose element of LOM,[1] will allow creating statements

[1]The values of the purpose element shall come from the following list of tokens: `discipline`, `idea`, `educational objective`, `accessibility restriction`, `educational level`, `skill level`,

about LOs like "prerequisite knowledge is boolean logic" or "educational objective" is control structures. The fact that the LOM standard allows developers to make choices about the data types makes it necessary to harmonize the metadata in distributed repositories before they can start communicating together. This procedure may be very costly.

Another important standard being used by providers of learning repositories and learning managements systems (LMS) is Shareable Content Object Reference Model (SCORM) [SCORM, 2009]. SCORM is not a completely new standard, but it is rather a "reference model" that integrates a set of inter-related technical standards (including LOM), specifications, and guidelines designed to meet requirements for learning content and learning systems.

SCORM explicitly supports the idea of learning content being aggregated into more complex units. Therefore, SCORM allows to specify a sequencing and navigation among learning objects in the form of rules which allow to define sequencing behavior independently of instructional content. On the other hand, it is essential to notice that SCORM sequencing does not address nor it precludes artificial intelligence-based computed sequencing. More specifically, the standard only allows to describe the flow and branching of learning resources based on an activity tree and an authored sequencing strategy. The activity tree is a conceptual structure of learning activities managed by the Learning Management System (LMS) and a learning activity references a resource that is presented to the learner. It is interesting to note that SCORM, as opposed to LOM, specifies how a LMS should interpret the data.

Overall, the standards give us vital information about what metadata is required by digital repositories and what types of applications can be built using this metadata. As we have seen, there is not a single standard that would be suitable for all digital repositories and applications and it is not probable that there will be one soon. The issue of interoperability, for example in case of LOM, is also not fully resolved. As a result, we think that work should focus on development of efficient metadata generation methods and tools rather than on development of an ideal metadata standard. If the metadata generation tools were in place, they would allow a digital repository to adopt a given standard easily and perhaps it would be possible to use even more standards at a time to meet the requirements of a variety of applications. The tools would also help to resolve the problem with updating of hardly maintainable metadata fields, such as relation fields.

So far, we were mainly concerned by the question of what metadata should be required by a metadata standard, however we did not explicitly specify the format in which this metadata information should be represented. An important step in this respect was made by three W3C recommendations: RDF [Klyne and Carroll, 2004], RDFS [Brickley and Guha, 2004] and OWL [Dean and Schreiber, 2004]. The recommendations standardize syntax for publishing machine readable data on the web. Since their release, many RDF(S)-based technologies have appeared that focus on the exploitation of semantic technologies for managing large data collections. To allow processing of large RDF data collections, the SPARQL query language [Prud'hommeaux and Seaborne, 2007] has been introduced and standardized by the W3C in 2008. Since then, further developments have been made in implementation of efficient triple stores (such as Sesame,

`security level`, `competency`.

Virtuoso or Josecki) capable of storing, indexing and retrieving large amounts of RDF(S) resources.

An important activity in this field is the development of Linked Data technology [Bizer et al., 2009]. Linked Data represents one of the first attempts to connect multiple large data sources on the web and to provide the means of exploring them through federated queries using semantic web technologies. Technically speaking, most of the data sources are sets of RDF(S) (or OWL) ontologies, or wrappers of those for standard relational databases, and SPARQL endpoints are being used to explore them. The initiative is open and anyone is allowed to connect their domain to the linked data federated repository. The main benefit this infrastructure brings is an open and easy-to-join architecture for connecting and querying data accessible through web technologies. Currently, the Linked Data network contains over seven billion RDF triples in domains as diverse as medicine, census data, and geographical data.

2.2 Automatic organisation of content

In the previous section, different approaches to interoperable representation of metadata were presented and the issue of how rich metadata description should be used for the representation of a resource was discussed. While finding a good format is in principle mainly the matter of an agreement, the decision on how rich description should be used is directly influenced by the tools that can help in the metadata generation process.

In this section, state-of-the-art methods that are not limited, but can be applied to automatic metadata generation are presented. We discuss approaches addressing all three types of metadata identified in chapter 1, i.e. keywords extraction (type 1 metadata), text classification (type 2 metadata) and links generation (type 3 metadata) to which most of the space will be dedicated. Issues related to the semantics of links and approaches to automatic generation of learning pathways will be also reviewed.

2.2.1 Keywords extraction

One of the most common types of metadata required by digital repositories are *keywords*. Keywords are entities describing the key concepts of a given resource. They are important for information retrieval of resources. Keyword extraction deals with methods that can automatically or semi-automatically extract keywords from the resource content.

Keywords are in practise usually specified for a resource in a free-text format or they are assigned to a resource in the form of a reference to an entity in a controlled vocabulary. This distinguishes two common approaches to keyword extraction from text.

- The *Automatic Term Recognition (ATR)* approach focuses on the extraction of words and multi-word expressions that are significant for a given domain. This approach employs a linguistic filter, based on part-of-speech (POS) tags, to extract a set of candidate terms. Term variant recognition techniques can be then applied to associate different realizations of one term with its root form. The candidate terms are then weighted using using various statistical measures.

- The *Thesauri/Ontology based* approach requires the set of possible candidate terms to be known in advance. The source text is then analyzed to find these terms and to associate them to concepts. The significance of a concept can be measured (if needed) in the same way as in the first approach.

The properties of the approaches will be now discussed. The shared characteristic of both approaches is that they will not discover a keyword if it does not appear in the resource text. However, we argue that keywords not appearing in the resource text should be rather treated as higher level topics/classes which can be discovered using the methods of section 2.2.2. As a result, extraction of keywords that appear in the resource text at least once are considered only.

Another characteristic is that both approaches are implicitly unable to deal with polysemy or homographs (words that share the same spelling but have different meanings). Fortunately, in well-established domains, such as in mathematics, terminology can be mostly considered unambiguous. However, it is important to keep in mind that its use in natural language can be more complicated. For example, words or collocations can sometimes play the role of keywords and sometimes not, such as in the case of the word *relation* referring either to the algebraic term n-ary relation or to any other kind of a relation, such as the relation between a dividend and a divider. Word sense disambiguation [Yarowsky, 1995] has in this respect the potential to improve the results of keyword extraction.

The roots of ATR dates back to late 1980s when the need for automatic extraction of terminological units from specialized texts became acute in various fields [Ahmad et al., 2005]. The advantage of ATR over the second approach is that there is no prior knowledge required, some of the statistical measures may only require a collection of general texts, such as the English Gigaword corpus. On the other hand, the second approach requires a robust vocabulary of terminology for a given domain. Such ontology may be freely available for certain domains, such as medicine, but may be more difficult to obtain for others. In terms of precision and recall, the first approach is likely to produce more *false positives* (a term which is in not a keyword was extracted), while the second one is likely to produce more *false negatives* (a term which is a keyword has been missed).

In terms of the quality of the final solution, the second approach has an advantage over the first one as ontologies usually contain information about the relatedness of terms, such as synonyms or hyponyms. As a result, both approaches can be combined in the following way. ATR is applied to find domain terms which are in turn used to induce a new ontology or update an existing one. The ontology may be then validated by domain experts to cut down the number of false positives. Later, the ontology is applied to keyword extraction.

Over the years, vast amount of unstructured data in digital format has encouraged researchers to employ ATR also for the tasks of glossary or index generation, tag suggestion and as the first step in automatic creation of thesauri and ontologies [Cimiano, 2006]. The results of ATR have also been successfully applied in information retrieval, machine translation and many other domains [Kozakov et al., 2004; Penas and Gonzalo, 2001; Dagan and Church, 1994]. In the meantime, many ontologies of domain terminology, such as Medical Subject Headings (MeSH) or NCI Thesaurus, have been applied to keyword extraction

from domain text.

Nowadays, there are quite a few ATR systems available via a web-based user interface [Sclano and Velardi, 2007] or as a web service [Yahoo!, 2008; Anadianou, 2008]. Some of them try to exploit the information provided by the annotation of particular formats. For example, Yahoo! Term Extraction Tool [Yahoo!, 2008] focuses on HTML documents and applies certain weight to particular HTML elements to determine what could be the most descriptive or significant terms. Others, such as [Sclano and Velardi, 2007], exploit the linguistic properties which makes them more applicable in general settings. Unfortunately, the access policy of the online tools often disallows researchers to experiment with the implemented methods and discourages advanced processing of the output to refine the results.

Despite its popularity, the ATR field still lacks proper comparative studies. Only a few methods have been evaluated and compared in terms of their precision. The rest of the developed tools is often assessed just by an observation, often concluding that it provides reasonably "good results". This issue is further discussed in the practical part of this report in section 4.1, more information can be also found in [Knoth et al., 2009].

2.2.2 Text classification

Text classification has been applied to many real world problems, such as to spam detection. Its importance grew quickly with the amount of information available on the web. Along with the widespread use of text classification methods comes the need for automated classification of new documents to hierarchies (for example, web directories). In this section, the state-of-the-art in hierarchical text classification will be discussed.

Hierarchies are becoming ever more popular for the organization of documents, particularly on the web. Hierarchies are very practical when a user wants to browse content according to its topic without having a specific goal in mind. The user can also easily see a bigger picture of the domain and it is easier to explore related domains than using keyword based searches. Large web directories available on the Internet, such as Open Directory Project or DMOZ, require lots of manual maintenance to validate the topic of newly submitted content and to filter out spam submissions.

There exists a great variety of methods for text classification. Many machine learning techniques were explored in the context of text classification including naïve Bayes, decision trees, k-Nearest Neighbour, Rochio, neural networks and support vector machines (SVMs). An overview of the text classification methods can be found in [Sebastiani, 2002; Manning et al., 2008]. One of the predominant approaches today is to convert documents into Vector Space Model (VSM) representation and to train machine learning classifiers on a set of labeled documents. The trained methods are then applied to an unseen set of documents. One of the common shortcomings of text classification methods is that they may require a great deal of supervision in terms of the amount of labeled documents. Techniques that are able to learn and generalize from a very small set (*seed*) of labeled documents by relying on pattern acquisition from an unlabeled set of documents are becoming very popular. These techniques are often denoted as *weakly-supervised* techniques. The most common variations of week supervision are *expansion* and *active learning*.

Classification problems can be divided to *one-of* and *any-of* problems. *One-of* problems assign a document into just one of N possible classes. *Any-of* problems assign a document into 0 to N classes. Both of these problems can be converted to binary decision problems that can only decide between two classes. Let us now describe the approaches that can be used for classifying into a hierarchy of classes. There are two common approaches capable of deciding whether a certain concept in the hierarchy corresponds to a given document:

1. Flattening the structure and building separate classifiers for each concept in the hierarchy. The final decision is based on a voting mechanism.

2. Building a separate classifier for each non-leaf node of a hierarchy which decides what branch should be followed from that particular concept.

Solution (1) is more general as it can deal with non-hierarchical structures as well. However, the classification phase is computationally expensive. To assign a class to a new instance, it involves all the classifiers built in the training phase. This solution can be used for both *one-of* and *any-of* problems. Solution (2) is more efficient in this respect (the average number of the involved classifiers is given by a logarithm of the number of classes in the ontology), but it is harder to apply it for *any-of* problems. Solution (1) can benefit from addressing each individual concept separately. When solution (2) is applied and the hierarchy contains many levels of classes, the information about the lower level concepts can be hard to access for the higher level classifiers (which have only very broad view on the distributions of the data in lower branches) [Grobelnik and Mladenic, 2005]. Nevertheless, solution (2) is usually preferred for large hierarchical structures due to efficiency reasons.

Related work on hierarchical classification of content into conceptual hierarchies can be found in [Frommholz, 2001; Cesa-Bianchi et al., 2006a; Dekel et al., 2004]. For example, experiments with SVM applied to hierarchical classification of web content were reported in [Frommholz, 2001]. The performance was tested on a two-level hierarchy which was created from a heterogeneous collection of pages. Somewhat surprisingly, it was found that there was no significant difference in accuracy between the two approaches. As a result of this, authors found solution (2) being a good choice, since it was approximately six times faster on the given dataset. A combination of the naïve Bayes algorithm and hierarchical SVMs which provided slightly better results on some testing data sets is described in [Cesa-Bianchi et al., 2006a]. A new on-line algorithm that incrementally learns a linear-threshold classifier for each node of the taxonomy is introduced in [Dekel et al., 2004].

In [Cesa-Bianchi et al., 2006b], the classification is also obtained in a topdown manner, so the algorithm is well-suited especially for the scenarios in which new data is produced frequently and in large amounts. To evaluate classifiers, the authors propose a multipath framework involving a new hierarchical loss function which captures the intuition that whenever a classification error is made on a node in the taxonomy, no loss should be charged for any additional error occurring in the subtree of that node.

So far, we mainly discussed classification into a structure that is hierarchical. However, there are also methods which can classify documents based on general ontologies [Tiun et al., 2001; He et al., 2004]. The basic steps of the system presented in [Tiun et al., 2001] can be summarized as follows:

- Identify concepts within the document. Map these concept to an ontology possibly enriched with concepts from an external knowledge base (e.g., Wordnet).
- Reduce the set of concepts to finally get only one concept which will label the document.

The authors conclude that although the method is quite simple, the results are comparable to those obtained using advanced ML methods.

2.2.3 Links and learning pathways

The notion of using links to facilitate the exploration and navigation over resources is relatively old. In 1945, Vannevar Bush published an influential article [Bush, 1945] where he considered a future device called "*memex*." According to Bush, memex allows an individual to store all their books, records and communications. Bush then goes in his ideas further, by establishing links as an essential part of memex and claiming that they correspond to a natural way how our mind operates:

> The human mind operates by association. With one item in grasp, it snaps instantly to the next that is suggested by the association of thoughts, in accordance with some intricate web of trails carried by the cells of brain.

In this way, Bush predicted the emergence of the hypertext where resources are linked together. Bush mentioned also trails, which can be understood as logically organized sequences of resources connected by links. However, while Bush saw links and the trails as being built manually by users, these days, as the amount of information stored on the web grows, we are starting to be more and more interested in automatic approaches.

In this section, the research in automatic generation of trails will be discussed. Automatic generation of links is then reviewed in section 2.2.4. We decided to introduce research in generation of trails before the automatic generation of links as it will help us to build the argument showing the need for robust link generation methods.

First, the work carried out so far in building trails from hypertext resources available on the Internet is presented. Later, we focus on automatic approaches to building trails in digital repositories used for learning. The term digital repository will be used to refer to a system into which independent resources can be submitted. Each resource is enriched with metadata as a necessary part of its submission procedure.

Creating trails in hypertext

When the Internet emerged, the Bush's ideas influenced many researchers. For example, the Walden's path system [Shipman et al., 1998] allowed to manually create trails - *learning pathways* by composing them from resources on the web. The authors suggested that the ordering of resources on the path should not be part of the learning content itself, but should be maintained on top of the content as a metadata structure. This is very similar to what is currently expected from

metadata standards, as described in section 2.1. Although positive outcomes of the so called "*pathcentric*" exploration on users were reported, the study did not address the infeasibility of this approach for larger collections and did not use any automatic approaches.

More sophisticated techniques for information sequencing in hypertext also appeared. For example, in [Guinan and Smeaton, 1992] a method for automatic generation of hypertext pathways in response to a user's query was proposed. The method searched for a set of resources relevant to a given query and then imposed a logical ordering on them based on the analysis of links shared by the retrieved resources. Unfortunately, the method analyses an existing manually created graph of links and the links are in addition required to be annotated by their type.

An important contribution to automatic construction of pathways in hypertext has been The Best Trail Algorithm published in [Wheeldon and Levene, 2003]. It is an iterative probabilistic algorithm based again on the analysis of a hyperlink graph. The Best Trail algorithm uses as an input a web graph, the set of possible starting nodes S, a parameter $M > 1$ which specifies the number of repetitions of the algorithm for each member of S and a *discrimination factor* (df) controlling the exploration of the algorithm. When the algorithm terminates it outputs a set of trails B. Another theoretical model for user navigation is presented in [Levene and Loizou, 2003] where an entropy of an underlying Markov chain modeling the web topology is measured. Each hypertext page is represented by a state and probabilities are associated with the links to different pages. The authors mention two interpretations of the probabilities: 1) The probability of following a link from a given resource by users 2) A score denoting the relevance users attach to a link given a particular query.

The methods for trails generation from hypertext typically rely on the analysis of the graph of resources. The edges of the graph representing links are in the current systems usually authored manually, however it has been shown in [Ellis et al., 1994] that there are significant differences in the links assigned by different people. Thus, the methods are unable to use an accurate model of the system.

Creating trails in digital repositories for learning

There are two main differences in the information used for the generation of trails in digital repositories for learning from the approaches presented in the previous section.

1. Metadata - static links among resources are in digital repositories expressed in different ways than in hypertext or may be even missing. On the other hand, additional metadata, such as classification information, would be typically available in digital repositories, but rarely available in the hypertext context

2. User model - the model of a user including properties, such as their competencies or their browsing history, may be available in the context of a digital repository.

An important task in digital repositories for learning is to recommend a user the most suitable resource or a logical sequence of resources in their specific

context. In order to achieve this, it is necessary that the system is aware of the content available, for which reason resources are annotated with metadata. In addition, as different users may have different needs, it is also necessary to know something about the user. Levene [Peterson and Levene, 2003] suggests that comparing these two may help to asses the suitability of a particular resource in a given context.

An approach which sees automatic composition of learning pathways as an optimization problem was presented by Knolmayer in [Knolmayer, 2003]. Knolmayer assumes that there is available an adjacency matrix of learning resources which in fact describes a graph of prerequisites. Each resource is annotated with additional information, such as the time required to study the resource or the quality of the resource as assigned by other users. The task is then to find the most optimal path through the graph according to a given criterion. While this approach already uses metadata which is typically not available in general hypertext, it does not explicitly work with the individual user's model.

An area of research dealing with adaption of systems with the goal to support their personal navigation experience is often referred to as *personalisation*. Such navigation can be performed in many different ways, see [Brusilovsky et al., 2007] for more information. The use of personalization approaches to automatic generation of learning pathways was reported in the SeLeNe project [Keenoy et al., 2004]. The authors developed a representation for definition of educational goals using the Bloom's taxonomy [Bloom et al., 1956] of educational objectives. Learning pathways were then derived by matching the learners objectives to metadata descriptions based on the relation element of the LOM standard.

As we know from section 1.1.2, it is very impractical to annotate a resource by information about its relation to other resources and therefore populating the relation element of LOM manually does not scale up. A possible solution to this problem is to develop an ontology of competencies and annotate each resource with the competencies required for its understanding. This approach has been followed, for example in [Kontopoulos et al., 2008], where a reasoning system was employed to plan the most suitable path through the available resources with respect to the current competencies of a learner. In this way, the problem is in fact translated into a classification problem which can be solved by applying techniques from section 2.2.2. A drawback of this approach is that the ontology of competencies and their relationship has to be specified manually. It has been found that this may be a very time consuming and tricky problem.

All of the presented approaches required a specification of ordering among resources either in the form of a static graph or a graph which can be created dynamically by matching a competence ontology with resources annotated by these competencies. Thus, none of the approaches was able to derive the ordering directly from the content. A fully automatic method for the calculation of a learning pathway based on the criterion of semantic similarity was published in our study [Zdrahal et al., 2009]. We used a hypothesis assuming that the prerequisite relation correlates with the criterion of semantic similarity and thus we searched for the most similar path in a $N \times N$ similarity matrix visiting each node/resource just once. Unfortunately, this leads to a well-known mathematical problem with high complexity called Traveling Salesman Problem (TSP).

After reviewing the state-of-the-art in automatic generation of trails, we indicated a few factors that we believe are the bottleneck for future progress.

First of all, the field is lacking a good evaluation model. One of the reasons for this may be that it is very subjective to decide what path is the most appropriate in a given context, another reason is probably the high $n!$ number of possible pathes for generating a course of length n. As a result, evaluation with respect to recall-like measures is very problematic. Thus, the possibility of evaluation is restricted to investigation of precision-like characteristics.

Another important bottleneck is that the currently used methods try to find the most suitable path, but they do not distinguish between the types of links they may follow. It is, in fact, essential to be able to distinguish whether the resources are related conceptually or because they have, for example, the same author. If they are only similar or one should be a prerequisite of the other. It would be naive to expect that high accuracy in automatic generation of trails can be achieved without this knowledge. Therefore there is a need for methods that are able to induce the graphs automatically and then assign a type to the links.

2.2.4 Automatic link generation and typing

The infeasibility of manual maintenance of links among resources in digital repositories was discussed in the introductory chapter. This creates an acute need for automatic or semi-automatic approaches to link generation. In this section, different approaches to link generation are first distinguished. Approaches that are applicable in our context are then reviewed. Later, the notion of link types and the existing methods for link typing are discussed.

Although there exists a great amount of scientific literature discussing link generation and relation extraction, the approaches often significantly differ both in the type of their input and the type of their output.

When considering the expected output, the approaches can be distinguished according to several criteria:

1. The type of linking with respect to a document collection (intra-document links, inter-document links)

2. The type of linking with respect to the granularity of the anchor and the link target (word/noun phrase, passage, file)

3. The semantic type of the generated link

An example of *intra-document* linking would be a task in which persons are linked to their jobs, based on a sentence, such as:
Alice works as a Sales Manager. Information Extraction is capable of discovering these types of relations from a natural language text [Knoth, 2008]. An example of *inter-document* linking is the generation of links among resources in a collection of documents based on their relatedness. The second type of the division is based on the anchor and target granularity. Different combinations are useful dependent on the task in hand. We provide here a few examples: A *passage-to-file* link (linking a quotation to its source document), *noun phrase-to-noun phrase* (linking a product to a company), *noun phrase-to-file* (linking a concept to its definition), *passage-to-passage* (linking a passage to its more detailed version) and other combinations. Finally, an important difference is

the semantic type of the generated links, as it will be discussed later in this section.

When considering the input of the approaches, methods can be divided to those relying on:

1. Structural characteristics
2. Graph analysis
3. Content-based analysis

The first group of systems uses structural elements of a resource, for example titles or internal tree structure of sections within a resource, following the hypothesis that documents with a similar structure are related. There exists a few approaches for structural similarity based, for example, on finding the least tree edit-distance [Zhang and Shasha, 1989] or using other measures [Augsten et al., 2005]. Other approaches, such as [Lakkaraju et al., 2008], combine structural and content-based properties. The second group of approaches require at least some links among resources to be given. The link graph can be then analysed to induce new links following the hypothesis that documents that are linked are related. Methods using links should be typically used in combination with content-based analysis as in [Adafre and de Rijke, 2005]. The third group of approaches - the content-based approaches analyse the textual content of a resource with [Green, 1999] or without [Allan, 1997] the use of background knowledge sources, such as ontologies, to detect resources that are conceptually related.

The variability of the link generation tasks makes it in general difficult to compare link generation methods as each combination of input and output behavior usually requires slightly different strategies. Perhaps the main achievements in link generation have been in *file-to-file* link generation [Huang et al., 2009] and on *noun phrase-to-noun phrase* linking (often referred to as relation extraction) [Katrenko, 2009].

File-to-file link generation

The *file-to-file* link generation can be seen as a traditional task addressed by the information retrieval field. Typical approaches find semantically related resources by calculating the semantic similarity of resources based on their document vectors [Allan, 1997; Green, 1998; Zeng and Bloniarz, 2004]. The document vectors are usually created by processing the text of the resources often applying techniques, such as tokenization, stop words filtering, stemming and weighting. A range of semantic similarity measures can be then applied to the calculation of similarity between the document vectors. Cosine, overlap, dice and Jaccard coefficients are widely used measures for the calculation of similarity $sim(\overrightarrow{x}, \overrightarrow{y})$ of the document vectors $\overrightarrow{x}$ and $\overrightarrow{y}$. If the calculated similarity is higher than a given threshold τ then a new link is generated.

Zeng [Zeng and Bloniarz, 2004] experimented with using manually and automatically extracted keywords for link generation with results indicating a substantial overlap between links generated based on manual and automatically extracted keywords. Thus it is possible to expect that not only keywords representing type 1 metadata, but also classes representing type 2 metadata may be used as features for link generation.

A survey on automatic link generation methods considering mostly the *file-to-file* scenario was published by Wilkinson and Smeaton in [Wilkinson and Smeaton, 1999]. One of the issues in that time, was to convert existing books into hypertext. As a result, a number of hypertext construction systems that have to deal with the link generation problem appeared in early 1990s. In the last decade, *file-to-file* generation of links based on semantic similarity became de-facto standard in large digital repositories, such as PubMed or ACM Digital Library.

Similarity methods have been also adopted by search engines. Google provides links to similar content and also uses similarity to detect pages that steel content from other pages to decrease their Page Rank. Although the similarity methods are widely used in practise, much work is still needed. For example, it is still not known how much is the quality of the generated links influenced by the use of background knowledge as reported by Green in [Green, 1999]. Green investigated the question whether links generated based on ontologies, using a lexical chaining method, are better than links generated based on simple term-repetition document vectors. It appeared that his results were not statistically significant and so it was not possible to support this hypothesis. Current work focuses on comparative evaluation of *file-to-file* link generation methods, cross-lingual link generation and link generation in very large data collections.

Noun phrase-to-noun phrase link generation

The *noun phrase-to-noun phrase* scenario has been seen as an important task mainly by the information extraction and ontology learning communities. The task usually performs deep analysis of text with the goal to recognize relationships between textual units focusing on relations expressed in a close proximity of each other, typically within a sentence. Information extraction techniques can be used to assign semantic roles, such as person or place, to the textual entities [Knoth et al., 2008]. Pattern matching techniques including machine learning can be then applied to detect various types of semantic relations [Rosario, 2005]. This is very interesting in the context of our study. It articulates the fact that relation/link typing has been so far much more researched in the context of relations between objects or entities than in the context of relations between larger textual units, such as passages.

There are several application areas of relation extraction. Among the most important ones are methods for automatic learning of ontologies and their population [Cimiano, 2006] from text. Traditional approaches for ontology learning have used manually created patterns, such as Hearst patterns [Hearst, 1998], the most recent research uses modern information extraction and machine learning techniques to find such patterns automatically. Formal Concept Analysis (FCA) [Wille, 2005] has been widely accepted as a mathematical foundation for ontology learning. Relation extraction techniques have been already applied in commercial applications, such as PowerSet. PowerSet analyses text of Wikipedia and parses it into subject-verb-object (SVO) triples. A question answering system is then used to find answers in a large collection of triples. One of the current research challenges is to construct from the triples an RDF graph which could be queried by semantic query languages, such as SPARQL. This is particularly difficult when the generated RDF should populate formal ontologies with

rich semantics using so called *ontology-based* information extraction methods [Welty and Murdock, 2006].

Other link generation approaches

In this report, we see *file-to-file* and *noun phrase-to-noun phrase* as the two most extreme cases of link generation. While the former is more concerned with finding related content, the latter follows the goal of creating a more formalized model of the analysed text. For the purposes of information retrieval, the former methods are vital for the exploration of content in large repositories and the latter methods are essential for being able to directly answer and reason about complicated queries. As indicated, both tasks received considerable attention by researchers. However, much less effort has been invested into approaches focusing on tasks lying in between of the extreme cases, i.e. problems, such as *file-to-passage*, *passage-to-passage* or *noun phrase-to-file*. These types of links are particularly suitable for providing very flexible navigation experience to a user of a digital repository as demonstrated in the motivating scenario (section 1.1.1).

Generating links pointing to units of a smaller granularity than a resource can be considered as a task of *passage* or *focused* retrieval, a subdomain of information retrieval where the search engine locates the relevant information inside the document instead of only providing a reference to the document. The standardized testing of search engines in this domain is preformed by the Initiative for the Evaluation of XML retrieval (INEX), which is with TREC, CLEF and NTCIR one of the four main Information Retrieval Evaluation Forums. INEX have recently become to play an essential role in the link generation task by defining tasks for evaluation of systems generating different links than only *file-to-file* links. One of the INEX evaluation tracks is Link-the-Wiki Track which includes a tasks to analyse the text of a resource and to recommend a set of incoming and outgoing links from an anchor text to the Best Entry Point (BEP) in other documents in the collection. This means that anchor text will be linked to a specific position in the target document - the best entry point for starting to read the referenced material from. Performing such evaluation is not an easy task and we will discuss it later in this section.

There is ample scope for text analysis methods to explore ways by which links at a granularity below whole document can be generated. Link generation methods in this category are often implicitly concerned with generation of links expressing a specific semantic relationship.

For instance, let us consider Wikipedia, which is a good example of a system having manually maintained links of the *noun phrase-file* type. Wikipedia places a relatively consistent policy on the semantic type of its links. Most of the links connect a representation of a concept (anchor) to its definition page. Automatic generation of these links has been one of the tasks at INEX Link-the-Wiki Track.

Another example may be given by Kolak [Kolak and Schilit, 2008] who provided a method for mining repeated word sequences (quotations) from very large text collections. The algorithm has been integrated with the Google Books archive and has allowed users navigating to popular passages across resources.

Link typing

Link typing approaches will be now reviewed. We will focus mainly on information retrieval and passage information retrieval approaches where lies the center of gravity of our motivation as described in section 1.1.1.

Perhaps one of the first people who thought about the importance of link typing was Randall Trigg [Trigg, 1983] who developed a taxonomy of link types already in 1980s. In his scenario, he expected the link typing to be performed manually. His link types carried very rich semantic information. For example, he envisaged *explanation*, *simplification/complication*, *continuation*, *critics*, *supported* and many other link types.

Although link typing was neglected by the first specification of Hypertext Markup Language (HTML), the specification now supports `rel` attribute which makes it possible to provide typed links using values, such as *Alternate*, *Start* or *Next*. However, the expressivity of the link types is much more structural rather than semantic as in Trigg's work. As a result, the `rel` attribute can hardly be used for advanced navigation or for some sort of a "smart" reasoning.

An influential article on automatic generation of typed links based on the textual content of resources has been published by James Allan in [Allan, 1996] and also in his dissertation [Allan, 1995]. Allan first classifies links into three categories:

1. *Pattern-matching links* - links that can be discovered automatically using simple techniques (for example, using very simple patterns)

2. *Manual links* - these links are the opposite opposite of the pattern-matching links and cannot be discovered automatically using current technology.

3. *Automatic links* are links, which cannot be discovered in a trivial way, but can be recognised using statistical techniques.

Allan then focuses on the automatic links. This category of links include, for example, a relationship in which one document expands a topic discussed in the other document or a so called "*tangent*" relationship in which the same topic is discussed from two different perspectives. The main contribution of Allan lies in development of a few heuristics that can be used for the discovery and typing of links and an algorithm which can be used for the calculation.

Allan's methods for automatic link generation discussed represent documents using VSM. All documents are analysed by splitting them into smaller parts, such as paragraphs or for example, using topic segmentation techniques [Reynar, 1998]. Similarity measures are then applied to all possible document pairs to recognise semantically similar segments and to generate links among them. This information is passed to a merging algorithm, which is used to consolidate links into a more simplified structure. Various hypotheses can be then applied to detect link types based on the pattern and the mutual position of the links.

Allan's research is perhaps the most important piece of work performed in link typing so far. The next challenges include evaluation, which can be considered as a necessary driving force for the improvement of the typing approaches. It is also important to notice that the automatic link typing methods proposed by Allan cover only a very restricted subset of the link types originally presented by Trigg. Another step is to incorporate the theoretical findings into real world

systems. Work in this area has been started by Blustein in [Blustein, 1999] who experimented with incorporating *definition*, *structural* and *similarity* links into journal articles. Unfortunately, all the link typing evaluations performed so far including Blustein's and Allan's used very small data samples. Therefore, it was difficult for the authors to generalize and to draw conclusions that would be supported by statistically significant results. Despite its importance, the automatic link typing field is at the moment still a relatively unexplored area.

Evaluation of link generation and typing approaches

An important feature of each information retrieval system is its ability to provide quantitative measure of its performance. Evaluating and comparing systems is not an easy task and it should not be underestimated. A traditional evaluation method is to compare the results of the system to a *golden standard*, which is often hand-crafted by a human annotator. Not only may this procedure be tedious, but on many tasks the annotations assigned by humans can hardly be considered correct. When evaluation with respect to a golden standard in a subjective task is performed, it is expected that a sufficiently high level of the so-called *inter-annotator* agreement was reached.

Although there exist many other criteria for judging the performance, information retrieval systems are usually evaluated using standard measures including mainly *precision* and *recall*. Definition and discussion of these measures can be found, for example in [Manning et al., 2008]. The main difficulty in evaluation of link generation systems in terms of these measures is the lack of a golden standard. Moreover, development of an annotated collection for link generation and link typing faces to the following problems:

- The number of possible links is very high even for small collections
- The inter-annotator agreement is in link generation low [Ellis et al., 1994]
- It is not known how the link generation tasks should be defined in terms of granularity for specific link types.

Comparative evaluation of link generation systems have been carried out in [Huang et al., 2008] where Wikipedia links were taken as a golden standard. The authors admit that as a result of the fact that Wikipedia links are not perfect (validity of existing links is sometimes questionable and useful links may be missing), comparative evaluation of methods and systems is still informative. It is naïve to expect that it will be possible to accurately measure *recall*-like characteristics on a very large collections soon. However, *precision* at top n where $n \in \{1, 2, \ldots, 100\}$ may be perhaps the most important characteristic for a majority of application.

There exist other possibilities for evaluation, such as performing a user-centered evaluation rather than evaluation using standard measures. For example, Blustein [Blustein, 1999] evaluated his system by measuring how much time users save when looking for a specific information with and without automatically generated links.

2.3 Defining the Gap

This chapter reviewed the state-of-the-art in the areas relevant to the motivating scenario described in chapter 1. This section provides a gap analysis which serves as a basis for the formulation of the research proposal in chapter 3.

In the beginning of this chapter, various standards for the annotation of resources were discussed. We learned that there are many factors which determine the usability of a metadata standard for a given digital repository. One of them refers to the fact that the richer metadata are required the more sophisticated applications that use the metadata can be developed. At the same time, certain types of metadata, such as metadata describing relations, may be difficult or infeasible to provide. This may prevent a digital repository from adopting a metadata standard or alternatively the digital repository may use a very simple description for its resources instead. As metadata standards mostly do not impose the ways or methods used to provide metadata, we believe that work should focus on the development of tools that facilitate the provision of certain metadata that is difficult to obtain manually in the first place.

Section 2.2 reviewed the state-of-the-art in methods that analyze the textual content of resources with the goal to extract various types of metadata. The review revealed that there is active research in this problem area carried out mainly by the information retrieval, natural language processing, machine learning and technology enhanced learning communities. The *keyword extraction* problem of section 2.2.1 revealed that there is a lack of comparative studies that would compare the various methods under the same conditions. The *text classification* methods of section 2.2.2 became quite a standard solution and they have been extensively applied to many real world problems on the web, such as spam detection. So far, the large-scale text classification research has focused on classification with large numbers of documents and/or large numbers of features, with a limited number of categories. The current research continues on large-scale hierarchical classification to category systems, such as DMOZ.

Despite its importance, the link generation fields described in section 2.2.4 still requires plenty of research. The field suffers from a lack of methods and comparative studies caused most probably by the lack of annotated datasets which are for this task extremely difficult to obtain. In comparison to traditional information retrieval tasks progress of which is annually evaluated, for example by the TREC conference, the link generation field evaluated by the INEX conference is relatively young and the scope of its evaluation is limited to Wiki systems. As digital repositories and information on the web are multilingual, a very demanding problem is currently how to perform metadata generation of all three metadata types so that it can be shared among different languages. We will address this issue in our pilot study in chapter 4.

Perhaps the most under researched area is the link typing field. While it seems that there is a common understanding that link typing is important, it is still not clear how it can be performed automatically in general settings. Moreover, as the problem may be slightly different for certain domains, it is not clear what link types are the most important to be distinguished. For example, in the domain of technology enhanced learning it seems that an important link according to the LOM standard would be the *prerequisite* link, however what about a *contrast* link? Although some link taxonomies exist, they are inconsistent. Another issue is that the current standards are usually taking into account

only links among resources, but are neglecting links among parts of resources. This is understandable as there is a lack of automatic methods that can provide such metadata and manual provision is infeasible. As a result, it is possible to conclude that the technology for link generation and typing should be the driving force and not the other way around. In other words, the standards may be updated if the technology is available. Otherwise, there is no need to perform such updates.

In section 2.2.3, the notion of links from a historical perspective and their usefulness in learning, in particular for the automatic course generation task, was presented. The need for methods that are able to automatically generate a course by organizing resources into a logical sequence has been acknowledged by many researchers. The following conclusions can be drawn from the existing work:

1. Any system working with large amounts of resources should rely on automatically generated links rather than on human specified links.

2. Only certain types of links are of real benefit to the reasoner which computes a learning pathway.

3. Both the available data and the user model of the learner should be taken into account by any automatic course generation reasoner. These points articulate the necessity for the link generation and typing methods without which it is hard to progress in the automatic course generation field.

To conclude, link generation and typing represent important, but still under researched problems. Progress in these tasks would have a direct impact on many application areas. Our research will therefore focus on link generation and link typing methods in the context of digital repositories including learning object repositories. To keep up with the current progress in the digital repositories field, our methods should be prepared to overcome the issues of multilinguality and should be integratable with multimodal repositories.

Chapter 3

Research Proposal

In this chapter, the research proposal is presented. The research questions being asked have emerged from the gap analysis (section 2.3) presented in the literature review. The research questions are first introduced in section 3.1. Section 3.2 then summarizes the expected areas of contribution. Later, the research approach is described. Finally, an expected progress plan is demonstrated in section 3.4.

3.1 Research Questions

The research questions are formulated in the following points:

- How can text analysis techniques support the annotation and the information retrieval in multilingual and multimodal digital repositories?
- How to automatically generate links between documents and document passages in large repositories?
- How can be the generated links typed?
- How much is the quality of link generation and typing influenced by features, such as keywords, background knowledge and ontologies?

3.2 Contribution

The expected contribution of this research is now presented:

- New methods for link generation.
- New methods for link typing important in the context of digital repositories and learning.
- An evaluation study of the link generation and typing methods.
- An integrated open-source framework of link generation and typing tools.
- A methodology for integration of the developed link generation and typing methods with multilingual and multimodal digital repositories.

To the best of our knowledge there was no study investigating link typing methods based on text analysis in the context of learning so far. In addition, a framework for the detection of many vital relations for learning, such as the *prerequisite* relation, using text analysis is not available. There are plenty of possible scenarios (section 1.1.1) in which link generation and typing can be valuable for digital repositories and for learning.

Please note that we do not plan to develop entirely new methods from scratch. We will rather build on existing approaches and foundations where applicable. The expected impact of this research proposal can be summarized in the following points:

- Impact on theory
  - The new link generation and typing methods will make it possible to enhance the navigation over resources and the maintenance of metadata in digital repositories. The methods will be applicable across many types of digital repositories.
  - The link typing methods will provide new input for a variety of reasoning tasks on the Semantic Web.
  - The link generation and typing methods will find its use in a number of information retrieval tasks, such as in the ranking of search results in digital repositories or in the detection of malicious sites.
  - The research outcomes will provide an important input for automatic course generation approaches.
- Impact on practise
  - New ways of integration of the link generation and typing methods into real-world systems will be researched.
  - Users of the Eurogene system, which will be developed as a case study in the context of this work, will be able to access and explore textual information in more flexible ways being aware of the up-to-date related content and avoiding the "lost in hyperspace" problem.

3.3 Current progress

The proposed research is being carried out in the context of the Eurogene project (The First Pan-European Learning Service in the Field of Genetics, Contract no. ECP-2006-EDU-410018) as anticipated in the PhD application proposal. The provides a real-world environment and a great amount of content for application and testing of the developed methods.

The following papers related to the topic of this report were refereed by an external body, were accepted and were/will be presented at conferences:

- Knoth, P., Collins, T., Sklavounou, E., and Zdrahal, Z. (2010) **EUROGENE: Multilingual Retrieval and Machine Translation applied to Human Genetics**, Demo at ECIR 2010, Milton Keynes, United Kingdom

Deliverable No.	Deliverable title
D 2.1	Requirement analysis and use scenarios
D 3.1	Initial sustainability plan with report on model service concept and potential user groups
D 4.1	Initial report on development of multilingual domain ontology
D 5.1	Multimedia content annotation: tools, procedures, and guidelines
D 5.3	Process, roles and responsibilities for quality assurance
D 6.1	Query and search facilities for educational packages and guidelines for their use
D 6.2	Multilingual query and search engine, tailored to EUROGENE
D 7.1	Integrated EUROGENE platform with user and maintenance documentation
D 10.1	Test protocol and evaluated user feedback
D 4.2	EUROGENE multilingual ontology and ontology-based services
D 6.3	Reasoning engine with user guidelines

Table 3.1: List of deliverables

- Knoth, P., Sova, J., and Zdrahal, Z. (2010) **Eurogene - The First Pan-European Learning Service in the Field of Genetics**, Znalosti (Knowledge) 2010, Jindrichuv Hradec, Czech Republic
- Knoth, P. (2009) **Semantic Annotation of Multilingual Learning Objects Based on a Domain Ontology**, Workshop: Doctoral consortium at EC-TEL 2009, Nice, France
- Zdrahal, Z., Knoth, P., Collins, T., and Mulholland, P. (2009) **Reasoning across Multilingual Learning Resources in Human Genetics**, ICL 2009, Villach, Austria
- Knoth, P., Schmidt, M., Smrz, P., and Zdrahal, Z. (2009) **Towards a Framework for Comparing Automatic Term Recognition Methods**, Znalosti (Knowledge) 2009, Brno, Czech Republic

The work was also presented at Open University's student conference in Milton Keynes based on the following paper:

- Knoth, P. **Automatic cross-lingual annotation of content based on domain ontology**, CRC PhD 2009, Milton Keynes

In addition, the work done so far included the design, and development of the Eurogene platform which is documented in the deliverables listed in Table 3.1. All the deliverables were accepted by the European Commission. Parts of the system that are related to the topic of this report will be presented in chapter 4.

The work done so far focused on building the Eurogene platform, gathering content and building metadata extraction tools.

- The Eurogene portal has been developed and deployed. The system has been made accessible to users at `http://eurogene.open.ac.uk`

- A multilingual collection of documents (currently about 2000 documents) were gathered and annotated with the support of the Eurogene tools. In addition, we have created in collaboration with the University of Sheffield a link generation and typing collection consisting of 30 resources. The collection can be used for evaluation of link typing and generation methods.
- We have build and integrated tools for annotation of resources. Annotation using all three metadata types as identified in section 1.1.2 is supported. Automatic methods have been applied to extraction of keywords and links.

3.4 Progress Plan

The previous section briefly summarized the actives undertaken in the last year of my PhD. The following work is planned in the future to allow answering the research questions:

- Experiments and comparative evaluation of link generation methods
- Further research in link typing methods
- Integration of the developed methods with the Eurogene platform

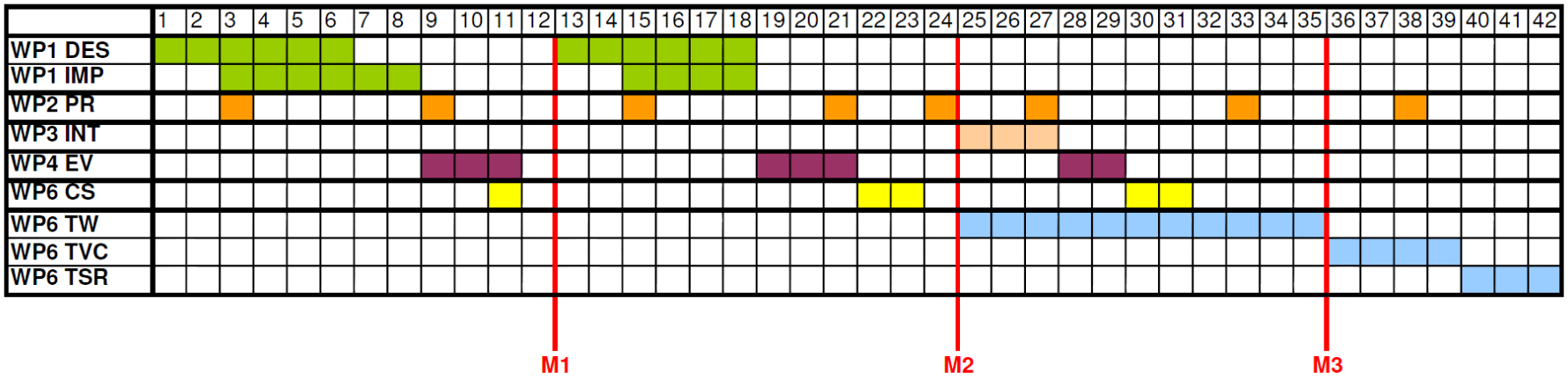
The results envisaged in section 3.2 will be achieved by disseminating the research outcomes at a number of relevant conferences. The following highly respectable conferences were identified as the most suitable international venues for disseminating the research results:

- Computational Linguistics (COLING) - every August
- European Conference on Information Retrieval (ECIR) - every March
- Special Interest Group on Information Retrieval (SIGIR) - every July
- European Conference on Technology Enhanced Learning (EC-TEL) - every September
- Language Resources and Evaluation Conference (LREC) - every even May

In addition, there are two venues directly targeting the research in the link generation field. These are:

- Workshop on Link Discovery: Issues, Approaches and Application (LinkKDD), which usually takes place at the International Joint Conference on Artificial Intelligence (IJCAI) - every odd July
- Initiative for the Evaluation of XML Retrieval (INEX) - every December

A detailed plan for the rest of this PhD project is now presented. The expected end date has been set to July 2013 which corresponds to a 4.5 years part-time PhD. This provides a one and half year contingency for the case of an occurrence of an unpredictable event. The work to be done has been split into tasks for each of which a time slot has been allocated. The plan is presented in the form of a Gantt chart in Figure 3.1.



WP1 DES - Design of link generation and typing methods
WP1 IMP - Implementation of the link generation and typing framework
WP2 PR - Progress reporting
WP3 INT - Integration
WP4 EV - Evaluation
WP5 CS - Conference submission and dissemination
WP6 TW - Thesis writing
WP6 TVC - Thesis validation and correction
WP6 TSR - Thesis submission and referreing

M1 - Milestone 1 - First evaluation study
M2 - Milestone 2 - Second year report
M3 - Milestone 3 - First thesis draft

Figure 3.1: Gantt chart

The planned work has been divided into six work packages (WPs) each of which consists of one or more tasks. WP1 is related to the writing of the thesis, WP2 corresponds to progress reporting, WP3 covers the design and implementation work, WP4 is used for integration, WP5 addresses evaluation and WP6 is intended for conference submissions. Three milestones have been planned. Milestone 1 will be achieved after first conference submission, Milestone 2 after the submission of the second year progress report and Milestone 3 with the first submitted draft of the dissertation. The PhD project ends in month 42 where month 1 begins in January 2010. It should be noted that the plan is informative and changes may have to be made, for example, to fit conference deadlines.

Chapter 4

Pilot Study

This chapter presents the pilot work carried out so far. The pilot study already directly addresses the first three research questions presented in section 3.1. In terms of link typing and generation, the chapter is structured along the ideas that both type 1 metadata and type 2 metadata as identified in section 1.1.2 may contribute to the generation of links (type 3 metadata). For example, the idea that keywords can contribute to generation of links was reported in [Zeng and Bloniarz, 2004]. It is therefore essential to have these techniques "in hand" for a possible use in the proposed link generation framework. The first two types of metadata are also a necessary prerequisite for integration of link generation methods into the Eurogene system, which serves us as a use-case scenario, and for being able to answer the first research question.

Section 4.1 presents a comparative evaluation of automatic term recognition techniques and a different approach that was applied to multilingual annotation with respect to the Eurogene ontology. Section 4.2 discusses the approach and the issues we have noticed while gathering and annotating content for the Eurogene system. Section 2.2.4 discusses the initial work carried out on the link generation and link typing tasks.

In order to gather sufficient mass of data that will allow us evaluation and integration of the proposed methods into a real-world system, it was necessary to build the Eurogene platform. Much of the time spent in the first year of my PhD was dedicated towards this goal. All the design and development work is documented in detail in the project deliverables listed in section 3.3. In addition, a general introduction to the functionalities of the portal is provided in our recent paper:

- Knoth, P., Sova, J., and Zdrahal, Z. (2010) **Eurogene - The First Pan-European Learning Service in the Field of Genetics**, Znalosti (Knowledge) 2010, Jindrichuv Hradec, Czech Republic

4.1 Keywords extraction and multilingual annotation

In section 4.1.1, various ATR methods were investigated and a comparative evaluation of the methods was performed on the Genia and on the Eurogene

collections. The study provided evidence that ATR methods may be very useful for identification of domain specific terms. In particular methods combining background and domain knowledge performed very well.

Section 4.1.2 deals with a different approach for keywords extraction based on domain ontologies. Theoretically, as suggested in section 2.2.1, the strengths of the both approaches can be combined. The presented method has been integrated with the Eurogene system and it can deal with annotation in multiple languages as addressed by the first research question. We also shortly outline how this scenario can be extended to multimodal repositories and show a cross-language information retrieval system we have developed and integrated with the Eurogene system.

4.1.1 Automatic term recognition

This section is based on the following paper:

- Knoth, P., Schmidt, M., Smrz, P., and Zdrahal, Z. (2009) **Towards a Framework for Comparing Automatic Term Recognition Methods**, Znalosti (Knowledge) 2009, Brno, Czech Republic

The section is organized as follows: We first outline the theoretical foundations of the methods implemented into a framework which originated as a library developed at the University of Sheffield and was extended in our work. More specifically, our contribution can be summarized as follows:

- Implementation of three ATR statistical methods (TF, RIDF and LR as described later in the text)
- Development of an automatic evaluation tool.
- Refactoring of the library (we had to fix quite a few bugs and added the possibility to choose a particular corpus as a background).

Later, the experimental evaluation on the GENIA and Eurogene corpora is presented. The section is concluded with the discussion of the possibilities for future work in this area.

Statistical ATR Methods

Let us recall that a typical approach of the advanced ATR methods consists of two phases:

- Linguistic phase employs a linguistic filter, based on part-of-speech (POS) tags, to extract a set of candidate terms. Term variant recognition techniques can be applied to associate different realizations of one term with its root form.
- Statistical phase uses a statistical method to assign a weight to each candidate term.

A concept may have many different surface realizations. For example 'human clones' and 'clones of human' could be considered as term variants. Identification of the term variants can have a positive impact on the results of ATR

	Termhood	Unithood
Only domain knowledge	TF, TFIDF, RIDF, DC	C-Value, LC
Background knowledge	Weirdness, LR, DR	

Table 4.1: Classification of statistical methods

methods [Nenadié et al., 2004]. Several types of term variations are usually distinguished – orthographic, morphological, structural, acronyms, abbreviations, lexical synonyms, etc.

To measure the 'strength' of a candidate term, two characteristics are usually distinguished – *termhood* and *unithood*:

- Termhood is a measure of the degree by which a linguistic unit is related to the domain-specific concept. Termhood methods are based on the frequency of occurrence [Kageura and Umino, 1996].
- Unithood is relevant for complex terms which consist of more linguistic units (words). It measures the collocation strength of the units. The basic idea of determining unithood consists in measuring significance of the words occurring together. Standard statistical techniques such as mutual information, t-test or log-likelihood are generally put to use [Ziqi Zhang and Ciravegna, 2008; Daille et al., 1994].

ATR methods can be also divided according to the use of background knowledge, i.e. a corpus in a general domain. Table 4.1 shows the classification of statistical measures that will be discussed in this section. Later, we will also discuss hybrid approaches that try to combine these measures. The following paragraphs briefly introduce particular ATR methods implemented in our framework that took part in the experimental evaluation reported in the next section.

Now we are ready to provide an overview of the state-of-the-art statistical methods for ATR that are implemented in the framework. As it was said, linguistic methods are usually language-dependent. To the contrary, statistical methods are usually language-independent, which explains our motivation in this work to put mainly statistical methods under scrutiny. The following paragraphs briefly introduce particular ATR methods implemented in our framework that took part in the experimental evaluation reported later.

Term Frequency (TF) is the count of all occurrences of the candidate term in a corpus. Frequent terms are supposed to be more important. This simple method is used in systems to rank term candidates generated by linguistic preprocessing [Dagan and Church, 1994]. We compute term frequency tf_i as a normalized frequency of term i in the document collection:

$$Tf(i) = \frac{f(i)}{\sum_k f(k)}$$

Term Frequency – Inverse Document Frequency (TFIDF) is a weighting score used often in information retrieval, where it corresponds to the fact

that the most significant words for a document tend to occur frequently in that document, despite their possibly rare occurrence in the whole collection. Inverse document frequency $Idf(i)$ measures the general importance of term i in the collection of documents D by counting the number of documents which contain term i:

$$\begin{aligned} Idf(i) &= log\frac{|D|}{|\{d_j : t_i \in d_j\}|} \\ TfIdf(i) &= Tf(i).Idf(i) \end{aligned}$$

Note that in the context of ATR we can prefer to compute a single ranked list of terms rather than a list of terms for each file in the domain-specific collection. Therefore, we can compute $tfidf(i)$ as $tf(i).idf(i)$ considering $tf(i)$ as the term frequency of word w_i in the domain collection. Roughly speaking, calculating the term frequency as there would be only one document in the domain-specific collection. The $tfidf(i)$ weighting score measures the termhood with respect to the documents in a collection. In ATR, it is often used as a baseline [Ziqi Zhang and Ciravegna, 2008] or as one of several features to determine the termhood [Medelyan and Witten, 2006].

Residual IDF (RIDF) is an alternative to IDF, which looks for terms whose document frequency is larger than chance. More precisely, RIDF is defined as the difference between logs of actual document frequency and document frequency predicted by Poisson distribution [Manning and Schütze, 1999].

$$RIDF(i) = Idf(i) - \log(1 - p(0; \lambda(i))),$$

where p is the Poisson distribution with parameter $\lambda(i) = \frac{f(i)}{D}$ (the average number of occurrences of word w_i per document). $f(i)$ is the number of words i in the collection. $1 - p(0; \lambda(i))$ is the Poisson probability of a document with at least one occurrence of i.

Weirdness measure is based on the idea that distribution of terms in a specialized corpus (domain) and in a general corpus (background) significantly differ [Ahmad et al., 2005]. This is expressed by the following formula:

$$Weirdness(i) = \frac{\frac{f_s(i)}{n_s}}{\frac{f_g(i)}{n_g}},$$

where $f_s(i)$ and $f_g(i)$ are the frequencies of word i in the specialized and the general corpus respectively, n_s and n_g are total numbers of words in the respective corpora. The original Weirdness was defined for one-word terms only, so we compute a geometric average of weirdnesses of each word in the term.

Likelihood Ratio (LR) [Manning and Schütze, 1999] is one of the methods we have newly implemented in the framework. The motivation is the same as in the case of weirdness. As opposed to weirdness, however, a statistical test is employed to measure the significance of difference between word frequencies in the domain and those in the background corpus. The first hypothesis is

that the probability of observing a given word in the background is equal to the probability of observing it in our domain. The second hypothesis is that the probability of seeing a given word in the domain is significantly higher than seeing it in the background. We assume binomial distribution for word frequencies.

$$p = \frac{f_s + f_g}{n_s + n_g} \qquad p_s = \frac{f_s}{n_s} \qquad p_g = \frac{f_g}{n_g}$$

$$LR = \log L(f_s, n_s, p) + \log L(f_g, n_g, p) - \log L(f_s, n_s, p_s) - \log L(f_g, n_g, p_g)$$

$$L(k, n, x) = x^k (1 - x)^{n-k}$$

Although Likelihood Ratio has been recently used in the related field of text summarization [Gupta et al., 2007], there is no quantitative evaluation of the method in the context of ATR to the best of our knowledge.

C-Value Method is a unithood method which has been used for term recognition in the medical domain, which typically contains a large number of complex terms [Frantzi et al., 2000]. The formula to compute it is based on three principles – extracting the most frequent terms, penalizing the nested terms that occur as a substring of a longer candidate term, and considering the length of the candidates (the number of the words they consist of):

$$\textit{C-value}(a) = \begin{cases} log_2|a| \cdot f(a) & \text{if } a \text{ is not nested} \\ log_2|a| \cdot (f(a) - \frac{1}{|T_a|} \sum_{b \in T_a} f(b)) & \text{otherwise} \end{cases}$$

where a and b are the candidate terms, f denotes the frequency and T_a is the set of candidate terms which contain a.

Glossex Method [Kozakov et al., 2004] is based on two heuristics. The first measure evaluates the degree of domain specificity (TD) which is equal to our definition of weirdness. The second measure investigates the idea of term cohesion. Let $|t| = n$ be the number of word forming term t. The term cohesion can be then expressed as:

$$TC_{D_i}(t) = \frac{n.tf_{t,D_i}.\log tf_{t,D_i}}{\sum_{j=0}^{n} tf_{w_j,D_i}},$$

where w_j is a j^{th} word in term t. The two measures are combined using two user adjustable coefficients α and β.

$$\textit{GlossEx}(t) = \alpha.TD(t) + \beta.TC(t)$$

```
<sentence><cons lex="IL-2_gene_expression" sem="G#other_name"><cons lex="IL-2_gene" se
m="G#DNA_domain_or_region">IL-2 gene</cons> expression</cons> and <cons lex="NF-kappa_
B_activation" sem="G#other_name"><cons lex="NF-kappa_B" sem="G#protein_molecule">NF-ka
ppa B</cons> activation</cons> through <cons lex="CD28" sem="G#protein_molecule">CD28<
/cons> requires reactive oxygen production by <cons lex="5-lipoxygenase" sem="G#protei
n_molecule">5-lipoxygenase</cons>.</sentence>
```

Figure 4.1: Example of a GENIA annotation file

Combining Statistical Methods It is often advantageous to combine several above-mentioned methods. For example, a mixture of entropy and log-likelihood ratio as measures of unithood and tf.idf characterizing the termhood has been explored in [Patry and Langlais, 2005]. Simple thresholds on each feature defined the weak classifiers, which were successfully combined by a kind of boosting algorithm. Similar combination of measures is discussed in [Vivaldi et al., 2001] in the context of term extraction from medical documents in Spanish.

Evaluation

As an example of the use of our evaluation framework, we present results of the experiments on two large annotated data sets – the GENIA and Eurogene corpora in this section.

Experiments on the GENIA Corpus GENIA corpus is a collection of biomedical documents that were compiled and annotated within the scope of the GENIA project [Collier et al., 1999]. The goal of the project was to develop text mining systems for the domain of molecular biology. The annotation process aimed at manual annotation terms in almost 2,000 MedLine abstracts.

Let us discuss the origin of two variants of the evaluation data set extracted from the GENIA corpus. Figure 4.1 shows an example of an annotated sentence from the corpus. It can be seen that both terms – *IL-2 gene expression* as well as the nested *IL-2 gene* – are considered valid. This approach can be beneficial for some tasks such as ontology building where the nested part of the term can often be interpreted as a hypernym of the complex term. On the other hand, the nested terms are not desirable in other situations as they can inflate the terminological glossaries and refer to general concepts rather than domain-specific ones. Considering the potential dichotomy, we prepared two versions of the "gold standard" list of GENIA terms. The first one contains all the annotated terms (including the nested ones), the second takes only the longest part as a term in the case of nesting.

In the linguistic pre-processing phase, we have extracted 32,521 candidate terms. This set was ranked by the statistical methods. We report the precision of the methods at 3 points (cuts): after first 20 highly ranked terms, after first 200 and after 2000 terms. Although the first may seem to be a very small sample for the evaluation, it is a relevant benchmark when considering ATR for keyword extraction or tag suggestion.

As in many other fields, ATR can benefit from combinations of the base methods employing various voting strategies. We have experimented with many different combinations and proved the potential boost in precision. Tables 4.2

No. of terms	TF	TFIDF	RIDF	LR	Weirdness	Glossex	C-Value	Vot. -TFIDF	Vot. LR -TFIDF	Vot. all
20	0,90	0,90	0,75	0,95	0,70	0,90	0,95	**1,00**	0,90	**1,00**
200	0,76	0,80	0,80	0,85	0,78	0,83	0,87	**0,96**	0,84	0,91
2000	0,70	0,71	0,70	0,63	0,64	0,62	0,67	**0,79**	0,67	0,73

Table 4.2: Precision on GENIA Corpus (nested terms)

No. of terms	TF	TFIDF	RIDF	LR	Weirdness	Glossex	C-Value	Vot. -TFIDF	Vot. LR -TFIDF	Vot. all
20	0,90	0,90	0,75	0,95	0,65	0,90	0,90	**1,00**	0,90	**1,00**
200	0,75	0,79	0,76	0,84	0,59	0,83	0,85	**0,94**	0,83	0,82
2000	0,67	**0,68**	0,63	0,52	0,47	0,61	0,59	0,67	0,58	0,60

Table 4.3: Precision on GENIA Corpus (without nested terms)

and 4.3 present the results of the base methods as well as the most promising combinations evaluated on the GENIA corpus with the English Gigaword Corpus as the background (for computing weirdness and other measures).

Considering only the base methods (not their combinations), the C-Value method and LR achieved very good results. This fact is surprising especially with respect to the success of the LR measure that is basically neglected by the ATR community. Another notable point is that the results achieved by TF, which is the simplest method, are not significantly worse than the results of TFIDF and that the method sometimes even outperformed RIDF.

The best performer showed to be the combination of Weirdness and TFIDF, which provided excellent results in both – nested and not-nested settings. The method combining all non-voting methods scored well, but still not as good as voted Weirdness-TFIDF.

As the size of the gold standard for the setting without nested terms is lower than that for the nested terms, it is natural that the values of the precision also decrease. However, the drop in precision is rather small for most of the methods on the first 200 terms. We suppose that the radically different pattern of weirdness in this respect has much to do with the characteristics of the background corpus. Nevertheless, this hypothesis needs to be verified in future work.

In order to inspect the impact of the background corpus size, we run our experiments in two other settings:

1. replacing the English Gigaword Corpus by the British National Corpus (BNC) which is by about one order of magnitude smaller than English Gigaword;

2. without any background data (labelled Null in the following table).

The results of this experiment are reported in Table 4.4. Only methods that use background are listed, other methods would produce the same results as reported in Table 4.2. All the experiments were performed in the nested settings. The best results for each corpus and method are highlighted.

No. of terms	LR	Weirdness	Glossex	Vot. Weirdness -TFIDF	Vot. LR -TFIDF	Vot. all
English Gigaword						
20	**0,95**	0,70	0,90	**1,00**	**0,95**	**1,00**
200	**0,89**	**0,78**	0,83	**0,96**	0,88	0,90
2000	0,65	0,64	0,62	**0,79**	0,69	**0,75**
BNC						
20	**0,95**	**0,80**	**0,95**	**1,00**	**0,95**	**1,00**
200	0,87	0,69	0,84	0,95	**0,89**	**0,92**
2000	0,62	0,63	0,61	0,80	0,68	0,72
MedLine						
20	**0,95**	0,75	0,75	0,95	**0,95**	**1,00**
200	**0,89**	0,71	0,67	0,89	0,88	**0,92**
2000	0,53	0,57	**0,65**	0,75	0,65	0,73
Null						
20	0,85	0,70	0,90	0,95	0,90	**1,00**
200	0,75	0,61	0,66	0,85	0,78	0,85
2000	**0,70**	0,49	0,50	0,66	**0,70**	0,66
English Gigaword + BNC						
20	**0,95**	0,75	0,90	**1,00**	**0,95**	**1,00**
200	**0,89**	0,77	**0,85**	**0,96**	0,88	0,91
2000	0,65	**0,67**	0,63	**0,79**	0,69	**0,75**

Table 4.4: Impact of different sizes of the background corpus

The results show that there is not a significant difference in using English Gigaword and BNC corpora. Even using both, one cannot expect significant improvements in precision. However, using no background knowledge significantly deteriorates the performance. Naturally, voting mechanisms are more robust since the fall of one method can be compensated by the other one.

Evaluation on the Eurogene Corpus The ATR methods have been also tested one the resources developed within the Eurogene project. So far, we have collected 210 presentations used mainly for teaching genetics at the university level. First, we converted the presentations into plain text. The size of the whole corpus is approximately 4 MB (600,000 words). The terms are not annotated in the texts so we asked domain experts to evaluate the results of the compared ATR methods.

During the linguistic phase, 34,617 candidate terms were extracted. They were ranked and sorted using each particular method. Then, we asked two experts from different branches of genetics to inspect first 100 terms produced by each method. Their task was to decide which terms are characteristic for the genetic domain.

The task may seem simple, but the domain experts found it ill-defined. The lack of a precise definition of "the characteristic domain term" showed to be the major problem. Some terms, such as *p-value*, are terms of a specific branch of genetics (here, statistical hypothesis testing). These terms were considered differently by statistical geneticist and by clinical or molecular geneticist. Also,

No. of terms	TFIDF	RIDF	LR	Weirdness	Glossex	C-Value	Vot. Weirdness -TFIDF	Vot. LR -TFIDF
100	0. 70	0.63	0.60	0.79	0.75	0.66	**0.98**	0.49

Table 4.5: Precision on Eurogene corpus

there were discussions on the terms found to be too general that were, finally, not accepted as proper terms (for example, *genetics*). The evaluators also found it difficult to be consistent across large set of results; the evaluation took considerable time.

The results of the experiment are reported in Table 4.5. As in the case of the GENIA corpus, we found that the method combining Weirdness with TFIDF provided the best precision. Other methods usually scored significantly lower. As these results were not expected, we asked the domain experts to assess the extracted terms from the qualitative point of view as well.

They found that the results of the Weirdness algorithm capture the important domain characteristics. At the same time, there were a few essential flaws in the output. Typical errors contained a name of an organization or a name of the author. This happens due to the absence of these terms in the general corpora and their high frequency in the domain-specific content. The high frequency of authors' names was due to the name re-occurring in the footer of each slide of their presentations. Such presentation style naturally results in generating noise for the statistical methods.

The TFIDF algorithm produced a list of terms which were probably characteristic for certain documents within the Eurogene corpus, but were often too general to be considered as domain-specific terms. We expect that this is caused by the fact that the TFIDF calculation does not involve any background knowledge.

The list of terms extracted using the combination of both the methods differs from those given by Weirdness and TFIDF separately. The extracted terms mainly consist of names of genes, substances and specific genetic terms. The combination produced significantly higher precision than the components.

Conclusions and Future Directions

The ATR evaluation framework discussed in this paper proved to be extremely useful for fast hypothesis formulation and testing. We have implemented new statistical ATR methods and compared their performance on the two included corpora. Many experiments have been also run with different combinations of statistical measures. The best results on both corpora were achieved by combination of Weirdness and TFIDF measures, which produced substantially better results than other methods.

The results were also inspected from the qualitative point of view. This leads to the conclusion that methods combining domain specific knowledge with background knowledge are generally more robust than methods using only one of these sources.

The results of our experiments are fully reproducible since all the source

codes, the data and the software for evaluation can be downloaded.[1]. We would like to encourage other researchers to contribute to the framework. It is especially important to add new evaluation data sets on which ATR techniques can be tested. The community could keep the set of statistical algorithms up-to-date as new approaches will arise. A web-based user interface will also be implemented in order to allow non-programmers to try and evaluate the system.

From the research point of view, we agree with [Ziqi Zhang and Ciravegna, 2008] that many of the items identified as terms fall into the category that Information Extraction (IE) traditionally extracts from texts. For example, names of genes, diseases, substances, methods, etc. The employment of the IE techniques including both – traditional machine learning and weakly-learning techniques (active learning, co-learning, or expansion) could significantly improve the precision. ATR and IE techniques can also co-operate. For example, the extraction of names of people and organizations is a typical task of IE. The result could be used to filter the list of candidate terms and thus to solve the problems mentioned in Section 4.1.1.

4.1.2 Multilingual annotation of content

This section describes a keyword extraction approach based on a multilingual domain ontology. The method has been implemented and integrated with the Eurogene system. The multilingual ontology is first briefly introduced. Later, the annotation process is described. The section is based on D 5.1 Eurogene deliverable.

Multilingual ontology

In the Eurogene project, we have developed an English monolingual domain ontology of genetics by merging 6 genetic glossaries[2] that contained a descriptive, but not too extensive, terminology for our domain. The terminology currently contains about 1,700 concepts. These concepts were translated by providers of educational content with the help of machine translation into 6 languages (English, French, Spanish, German, Italian and Lithuanian). The providers were instructed to provide all possible versions (terms) of a concept being used in their target language. The ontology currently contains more than 12,000 terms.

The terminology is represented in a Simple Knowledge Organization System (SKOS) like structure. Using SKOS, concepts can be easily labeled with lexical strings in one or more natural languages. In particular, SKOS defines for a resource property `skos:prefLabel` and `skos:altLabel`. The former can be used to specify a preferred string label for a concept in a particular language while the later is used to specify an alternative string label for a concept. In this way, SKOS helps us to connect different representations of the same concept in multiple languages. SKOS also allows to specify relations between concepts, such as `skos:broader`, `skos:narrower` and `skos:related`, that are used to create *isa* hierarchies and to refer to related concepts in a vocabulary.

[1] `http://code.google.com/p/jajatr/`

[2] Published by the University of Washington in Seattle, National Institute of General Medical Sciences in Bethesda, Elsevier, Oracle ThinkQuest, University of Michigan and Centre for Genetics Education in Sydney

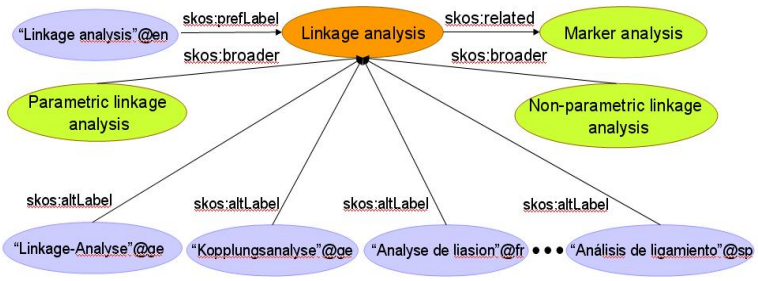


Figure 4.2: Representation of a concept *linkage analysis* in the multilingual ontology

Figure 4.2 shows how a genetic concept *linkage analysis* is represented in our ontology. The preferred label of this concept is the English version *Linkage analysis*. The concept has a two alternative representations in German (*Linkage-Analyse* and *Kopplungsanalyse*). The representation in French is *Analyse de liasion* and in Spanish *Análisis de ligamiento*. The concept *Linkage analysis* is a broader concept for *Parametric linkage analysis* and *Non-parametric linkage analysis*, and it is related to a concept *Marker analysis*.

More information about the Eurogene ontology can be found in D 4.1 and D 4.2 Eurogene deliverables.

The annotation process

The ontology of genetic terms can be easily used to annotate resources. In order to make this process as easy as possible for the content providers, our system allows them to upload their learning objects in different formats including `.doc`, `.ppt` and `.pdf`. The content provider first uploads a resource which is then stored in a database and converted to text. The source language of the LO is either specified by the content provider in advance or it is automatically recognized by the system. Words in the text are then converted to their root form (lemma) by applying stemming [Porter, 2000]. When the stemming is finished, the terminology of the detected language is loaded and applied to find all of the occurrences of the terms present in the source text.

In order to make the indexing process as fast as possible, the terminology is stored in the *trie* datastructure in main memory. The trie datastructure is an n-ary tree where edges correspond to letters and leave nodes to words of the terminology. Checking whether a word w is present within the trie datastructure takes at maximum $|w|$ steps where $|w|$ corresponds to the number of letters that compose word w. Therefore, the time complexity is constant with respect to the number of words in the ontology. The result of the annotation process is a set of domain specific terms with their term frequency that are present in a given learning object.

As the ontology connects different syntactic representations of a concept, it allows us to abstract to a language independent representation, i.e. from terms to concepts. This means that each resource may be represented using a Vector Space Model (VSM) where dimensions of the vector correspond to concepts.

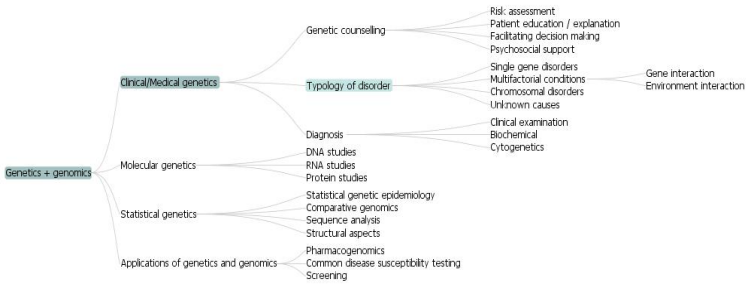


Figure 4.3: Part of the genetic hierarchy. The hierarchy can be interactively visualized and used for navigation.

More specifically, each resource is represented by a vector of length n, where n is the number of concepts in the multilingual domain ontology. Non-zero values of the vector correspond to concepts acquired by abstracting from terms found in the resource's text.

4.2 Lessons learned in content gathering and classification

This section provides information about the development of a classification hierarchy in the context of the Eurogene project. Our approach and current issues are presented.

4.2.1 Topic classification

The theme/topic hierarchy shown in Figure 4.3 was developed in close interaction with domain experts. In standard knowledge acquisition sessions, groups of domain experts were interviewed, their responses evaluated and the theme hierarchy constructed. In the next session, the results were presented back to the group for comments and corrections. This was repeated until a stable result was achieved. When all trees for all subdomains were completed, they were merged together and duplicity discussed and resolved. The proposed tree was eventually validated by the rest of the community.

The purpose of this hierarchy is to provide a coarse-grain description of the domain that can be used both to speed up search and to prune the search space used for more complex queries. Providing such a speed-up may also help link generation methods to significantly reduce their processing time. The maximum depth of the hierarchy is ten. At present, a resource is associated with one or more topics from the hierarchy manually.

4.2.2 Lessons learned

A few important remarks were noticed during the topic hierarchy creation, topic annotation and the topic hierarchy maintenance.

- It was very difficult to reach a certain level of an agreement between domain experts in the hierarchy creation process. The last version of the topic hierarchy is not the best possible, but rather a version which is acceptable across the community.
- The manual classification with respect to the hierarchy takes a significant amount of effort and requires a good level of domain knowledge.
- Annotations produced by inexperienced users are often inconsistent with annotations provided by experienced users.
- Online updates of the hierarchy are extremely costly as they may require re-annotation of a certain amount of resources. A practical problem which often appears then is how to ask a content provider to redo his/her annotation, because the hierarchy changed.

Performing the pilot work in this area helped us to reveal that tools that would at least assist in the classification process would be of real value. The maintenance of type 2 metadata as identified in section 1.1.2 is without such tools very difficult.

4.3 Cross-language discovery of related content

This section presents the work done so far on the automatic discovery and typing of links. Section 4.3.1 describes an approach for cross-language discovery of links of the *file-to-file* type. The approach was integrated with the Eurogene system where it is being used.

Section 4.3.2 describes a very experimental implementation of similar methods as suggested by [Allan, 1995]. So far, we have experimented with passage similarities based calculated using keywords. An annotated collection of resources taken from the Eurogene repository has been created specifically for this reason. The development of these methods is work in progress.

4.3.1 Link discovery

The information provided in this section is based on the paper:

- Zdrahal, Z., Knoth, P., Collins, T., and Mulholland, P. (2009) **Reasoning across Multilingual Learning Resources in Human Genetics**, ICL 2009, Villach, Austria

and the D 6.3 Eurogene deliverable as listed in section 3.3.

The Eurogene multilingual domain ontology includes the terms used in the supported languages to denote each concept. As reported in section 4.1.2, these terms can be matched against the text contained in a contributed resource in order to automatically suggest possible concepts that can be used for annotation. Validated concepts can be used for identifying semantically similar resources in different languages by comparing their associated (language independent) concepts.

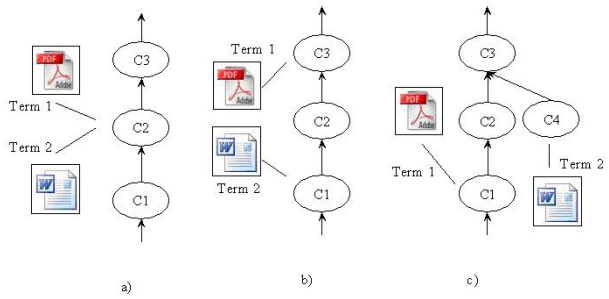


Figure 4.4: Terms describing the same concept (a) and two different concepts (b, c)

The annotation of resources makes it possible for algorithms to combine the concept matching based on semantic similarity with traditional string matching. We define three different measures for similarity based concept matching:

- **Identity** - verbal expressions (term) of two different educational resources are considered equal if the corresponding concepts are identical, i.e. if they belong to the same synset defined across all supported languages.
- **Generalisation** - for concept matching, two concepts count as equal, if one is a generalisation of the other and their distance in the hierarchy is shorter that a predefined threshold parameter, and
- **Common hypernym** - for concept matching, two concepts count as equal, if they share a common hypernym and their edge distance in the tree is shorter than a predefined parameter.

These three cases of concept matching are shown in Figure 4.4. Term 1 and Term 2 can be in different languages. If the distance parameter is greater than 3, then concepts C1, C3 in b), but not C1 , C4 in c) count as equal. If its value is 4 then the concepts both in b) and c) are considered as equal. Applying measures according to b) and c) requires additional heuristics to resolve the situations where the same concept might be included more that once. In the rest of the paper we will use only the first measure of concept matching.

One of the possible measures for calculating semantic similarity is the *cosine similarity* measure. The cosine similarity between two resources represented by concept vectors A and B is computed as follows:

$$cos(\theta) = \frac{\overrightarrow{A}.\overrightarrow{B}}{|\overrightarrow{A}||\overrightarrow{B}|}$$

Within the portal a resource similarity block has been implemented that lists similar resources to a displayed resource (see Figure 4.5). If the calculated similarity is higher than a given threshold $\tau = 0.4$ a similarity link is generated.

The calculation module uses a binary vector to represent each resource where the occurrence or absence of each of the domain concepts is indicated using values of 1 or 0 (respectively). This enables the comparison of resources across

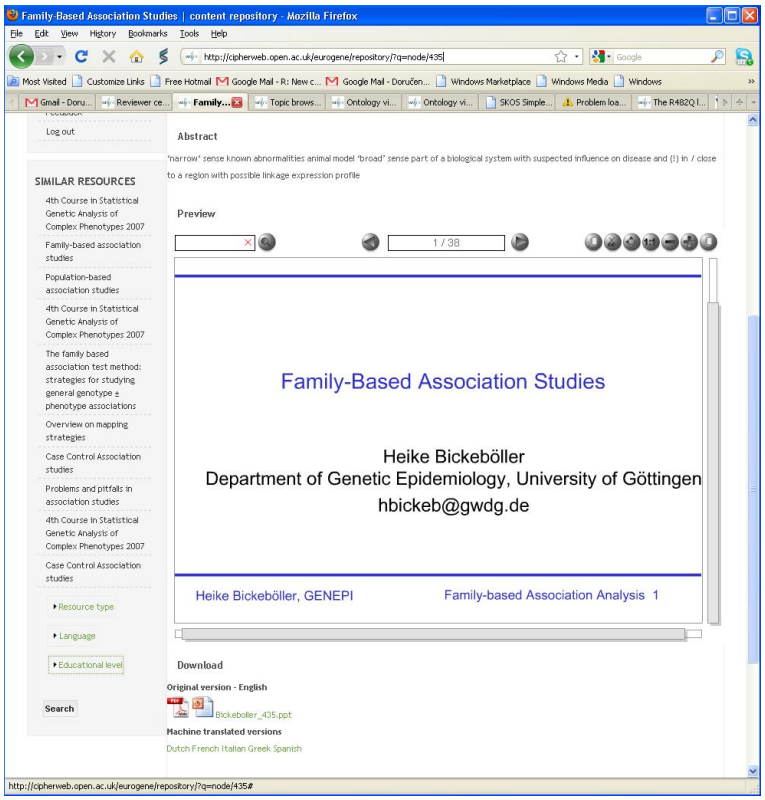


Figure 4.5: A block showing similar resources integrated with the Eurogene system on the left of the page

different media. Although the relative concept frequency can be a useful indicator of importance with text-based resources this cannot be attributed in the same way to images or video resources. By identifying conceptual similarity through concept occurrence it is possible to compare resources consistently across different media types.

4.3.2 Link typing

This section has been initiated by the paper where a few hypothesis for link typing inspired by [Allan, 1995] were discussed:

- Knoth, P. (2009) **Semantic Annotation of Multilingual Learning Objects Based on a Domain Ontology**, Workshop: Doctoral consortium at EC-TEL 2009, Nice, France

Dataset

After the paper was published, a dataset including relations among 30 resources was created. The resources were pdf or powerpoint presentations from various authors used for teaching statistical genetics. A domain expert performed the selection and annotation of the resources. All possible pairs were investigated. If two resources were decided to be related, a relation type was selected from one of the following choices:

- *Summarizes* - the first resource briefly covers a topic which is in more depth discussed in the second resource.
- *Prerequisite* - the relation describes that a user should read the first resource before the second one.
- *Different perspectives* - the first resource is similar to the second one, but it addresses the problem from a different perspective.
- *Similar narrative* - the first resource is semantically similar to the second one and in addition both use a similar narrative.
- *Similar* - the first resource is similar to the other, but the relationship is not specified.

Method

The task was to automatically determine the relationship between two resource given their content. The resources were automatically split into parts (individual slides) and the cosine measure was used to calculate similarities between the parts of two documents. If the similarity between the parts was higher than a manually set threshold a link between the parts was induced. All the relationships were then visualized. Our goal was then to experimentally test whether the relation between two resources can be determined based on the resulting pattern.

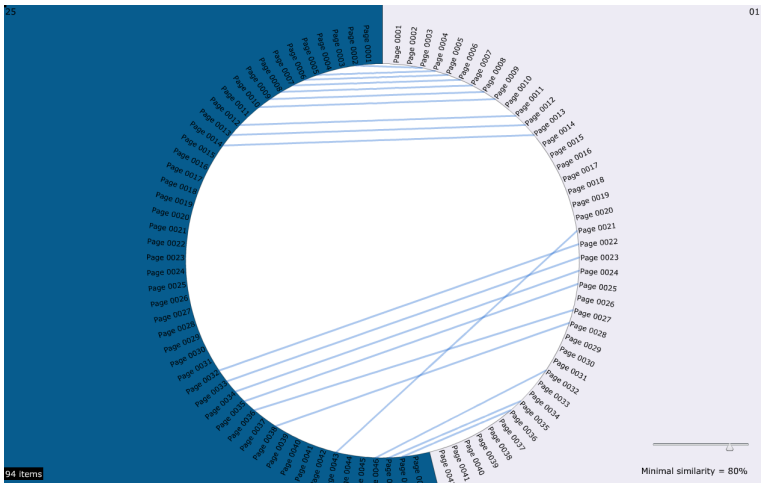


Figure 4.6: Similarity diagram. Parts of the resources follow the same narrative

Results

The whole procedure described in section 4.3.2 was performed by a tool which we have implemented for this purpose and which will be further enhanced during the PhD. The development of the methods for link typing is still in progress and therefore we will not present in this section any empirical results. However, we will show a few interesting examples and will speculate about the used approach.

For simplicity, the visualizations of the patterns were also manually investigated, assuming that if a computer programm should be able to determine the correct relationship based on the visualized pattern, a human should be also able to do this. We have experimented with different features (keywords extracted using the Eurogene ontology vs. simple term frequencies) and with different value of the similarity threshold τ.

A few examples of the diagrams are now presented. Figure 4.6 shows an example of a relation where parts of the two resources follow the same narrative. Since the similarity threshold τ was set to a relatively high value 0.8 and the pattern where links are not crossing was still visible, the parts of the document were investigated manually. It turned out that the first lecture used only negligibly modified slides from the second lecture.

Figure 4.7 presents a relation describing that one lecture summarizes the other. It has been reported in [Allan, 1995] that this type of relationship can be determined by a *V-shape* pattern.

The relationship in Figure 4.8 was annotated by the domain specialist as a prerequisite relationship. It is interesting to see that the shape is very similar to the summarize relationship. This shows us that the distribution of terms in a prerequisite relationship may be very similar to the distribution of terms in a summarize relationship, i.e. there may not be a simple way to distinguish between the case in which some terms were theoretically introduced in the first resource and then were used in the second one and the case where some terms were just accumulated into a shorter resource for the reason of providing an overview.

Figure 4.7: Similarity diagram. The light blue resource summarizes the dark blue resource

We believe that the features that would allow to distinguish these relation types are at the semantic rather than a syntactic level. More specifically, information about the prerequisite relationship between individual terms may be needed to be able to determine the relationship at the level of a document or a part of a document.

A couple of other interesting points were noticed during the experiments. We found that the optimal value of the similarity threshold is very dependent on the features used and the task in hand. For example, when similarities were calculated based on keywords extracted using the Eurogene ontology, the value of τ at which it was possible to see the pattern was in general higher than when simple term frequency features extracted from text were used. This can be explained by a substantially lower number of indexed terms in the former case than in the latter. In many cases the slide similarities were calculated based on a very low number of keywords (usually between 0-3 keywords). This can be explained by two main reasons: 1) Many keywords are missed because the ontology is incomplete 2) The text of a slide is very short and does not contain many keywords. As a result, the questions what is the optimal unit for splitting and how to index keywords that are not present in the ontology should be addressed in the future.

Overall, we have shown that link typing methods are capable of producing interesting results, but further research is needed to address many issues including the one formulated in the last research question (section 3.1).

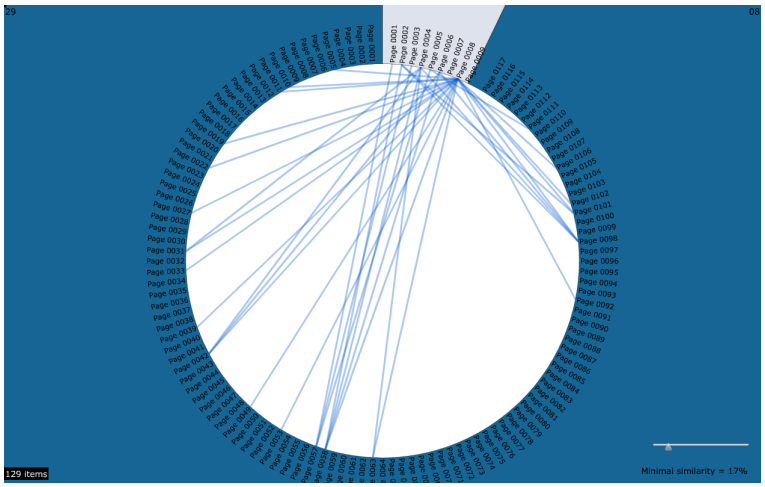


Figure 4.8: Similarity diagram. The light blue resource is a prerequisite for the dark blue resource

Chapter 5

Summary

This report dealt with the role of metadata in digital repositories. The types of metadata that are difficult to maintain were identified and the need for automatic approaches was articulated. State-of-the-art methods and approaches were then reviewed, which resulted in the formulation of the research questions.

It was found that in particular automatic content-based approaches to link generation and typing are currently very much needed as the content linking task can hardly be performed by humans. While it can be expected that progress would have a direct impact on a number of application domains, both link generation and link typing fields are still relatively unexplored.

As a result, we have formulated the research questions along these ideas and presented a research plan which will help us to answer them and to disseminate the research outcomes.

Bibliography

Adafre, S. F. and de Rijke, M. (2005). Discovering missing links in wikipedia. In *LinkKDD 05: Proceedings of the 3rd international workshop on Link discovery*, pages 90–97, New York, NY, USA. ACM.

Ahmad, K., Gillam, L., and Tostevin, L. (2005). University of Surrey participation in TREC 8: Weirdness indexing for logical document extrapolation and retrieval (WILDER).

Allan, J. (1995). *Automatic Hypertext Construction*. PhD thesis.

Allan, J. (1996). Automatic hypertext link typing. In *HYPERTEXT '96: Proceedings of the the seventh ACM conference on Hypertext*, pages 42–52, New York, NY, USA. ACM.

Allan, J. (1997). Building hypertext using information retrieval. *Inf. Process. Manage.*, 33:145–159.

Anadianou, S. (2008). Termine.

Augsten, N., Böhlen, M., and Gamper, J. (2005). Approximate matching of hierarchical data using pq-grams. In *VLDB '05: Proceedings of the 31st international conference on Very large data bases*, pages 301–312. VLDB Endowment.

Bateman, S., Brooks, C., Mccalla, G., and Brusilovsky, P. (2007). Applying collaborative tagging to e-learning. In *Proceedings of the 16th International World Wide Web Conference (WWW2007)*.

Bizer, C., Heath, T., and Berners-Lee, T. (2009). Linked data - the story so far. *International Journal on Semantic Web and Information Systems*, 5(3):1–22.

Bloom, B. S., Engelhart, M. B., Furst, E. J., Hill, W. H., and Krathwohl, D. R. (1956). *Taxonomy of educational objectives. The classification of educational goals. Handbook 1: Cognitive domain*. Longmans Green.

Blustein, W. J. (1999). *Hypertext Versions of Journal Articles: Computer Aided linking and realistic human-based evaluation*. PhD thesis.

Brickley, D. and Guha, R. V. (2004). Rdf vocabulary description language 1.0: Rdf schema. W3c recommendation, W3C. http://www.w3.org/TR/2004/REC-rdf-schema-20040210/.

Brooks, C. and Mccalla, G. (2006). Towards flexible learning object metadata. In *International Journal of Continuing Engineering and Lifelong Learning*, pages 50–63.

Brusilovsky, P., Kobsa, A., and Nejdl, W. (2007). *The Adaptive Web*. Springer.

Bush, V. (1945). As we may think. *The Atlantic Monthly*.

Cesa-Bianchi, N., Gentile, C., and Zaniboni, L. (2006a). Hierarchical classification: combining bayes with svm. In *ICML '06: Proceedings of the 23rd international conference on Machine learning*, pages 177–184, New York, NY, USA. ACM.

Cesa-Bianchi, N., Gentile, C., and Zaniboni, L. (2006b). Incremental algorithms for hierarchical classification. *J. Mach. Learn. Res.*, 7:31–54.

Cimiano, P. (2006). *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Collier, N., Park, H. S., Ogata, N., Tateishi, Y., Nobata, C., Ohta, T., Sekimizu, T., Imai, H., Ibushi, K., and ichi Tsujii, J. (1999). The genia project: corpus-based knowledge acquisition and information extraction from genome research papers. In *In Ninth Conference of the European Chapter of the Association for Computational Linguistics (EACL-99*, pages 271–272. http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA.

Cress, U., Dimitrova, V., and Specht, M., editors (2009). *Learning in the Synergy of Multiple Disciplines, 4th European Conference on Technology Enhanced Learning, EC-TEL 2009, Nice, France, September 29 - October 2, 2009, Proceedings*, volume 5794 of *Lecture Notes in Computer Science*. Springer.

Dagan, I. and Church, K. (1994). Termight: identifying and translating technical terminology. In *Proceedings of the fourth conference on Applied natural language processing*, pages 34–40, Morristown, NJ, USA. Association for Computational Linguistics.

Daille, B., Éric Gaussier, and Langé, J.-M. (1994). Towards automatic extraction of monolingual and bilingual terminology. In *Proceedings of the 15th conference on Computational linguistics*, pages 515–521, Morristown, NJ, USA. Association for Computational Linguistics.

Dean, M. and Schreiber, G. (2004). OWL web ontology language reference. W3C recommendation, W3C.

Dekel, O., Keshet, J., and Singer, Y. (2004). Large margin hierarchical classification. In *ICML '04: Proceedings of the twenty-first international conference on Machine learning*, page 27, New York, NY, USA. ACM.

Ellis, D., Furner-Hines, J., and Willett, P. (1994). On the measurement of inter-linker consistency and retrieval effectiveness in hypertext databases. In *SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 51–60, New York, NY, USA. Springer-Verlag New York, Inc.

Frantzi, K., Ananiadou, S., and Mima, H. (2000). Automatic recognition of multi-word terms: The C-value/NC-value method. *International Journal on Digital Libraries*, V3(2):115–130.

Frommholz, I. (2001). Categorizing web documents in hierarchical catalogues. In *In Proceedings of ECIR-01, 23rd European Colloquium on Information Retrieval Research*.

Ganter, B., Stumme, G., and Wille, R., editors (2005). *Formal Concept Analysis, Foundations and Applications*, volume 3626 of *Lecture Notes in Computer Science*. Springer.

Green, S. J. (1998). Automated link generation: can we do better than term repetition? *Comput. Netw. ISDN Syst.*, 30(1-7):75–84.

Green, S. J. (1999). Building hypertext links by computing semantic similarity. *IEEE Trans. on Knowl. and Data Eng.*, 11(5):713–730.

Grobelnik, M. and Mladenic, D. (2005). Simple classification into large topic ontology of web documents. *CIT*, 13(4):279–285.

Guinan, C. and Smeaton, A. F. (1992). Information retrieval from hypertext using dynamically planned guided tours. In *ECHT '92: Proceedings of the ACM conference on Hypertext*, pages 122–130, New York, NY, USA. ACM.

Gupta, S., Nenkova, A., and Jurafsky, D. (2007). Measuring importance and query relevance in topic-focused multi-document summarization. In *ACL*. The Association for Computer Linguistics.

He, Q., Qiu, L., Zhao, G., and Wang, S. (2004). Text categorization based on domain ontology. In Zhou, X., Su, S. Y. W., Papazoglou, M. P., Orlowska, M. E., and Jeffery, K. G., editors, *WISE*, volume 3306 of *Lecture Notes in Computer Science*, pages 319–324. Springer.

Hearst, M. A. (1998). Automated discovery of wordnet relations. In *C. Fellbaum, WordNet: An Electronic Lexical Database*, pages 131–153. MIT Press.

Huang, W. C., Geva, S., and Trotman, A. (2009). Overview of the inex 2009 link the wiki track.

Huang, W. C., Trotman, A., and Geva, S. (2008). Experiments and evaluation of link discovery in the wikipedia.

Kageura, K. and Umino, B. (1996). Methods of automatic term recognition: A review. *Terminology*, 3(2):259–289.

Katrenko, S. (2009). *A Closer Look at Learning Relations from Text*. PhD thesis.

Keenoy, K., Levene, M., and Peterson, D. (2004). Personalisation and trails in self e-learning networks.

Kittur, A., Suh, B., Pendleton, B. A., and Chi, E. H. (2007). He says, she says: conflict and coordination in wikipedia. In *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 453–462, New York, NY, USA. ACM.

Klyne, G. and Carroll, J. J. (2004). Resource description framework (RDF): Concepts and abstract syntax. World Wide Web Consortium, Recommendation REC-rdf-concepts-20040210.

Knolmayer, G. F. (2003). Decision support models for composing and navigating through e-learning objects. In *HICSS '03: Proceedings of the 36th Annual Hawaii International Conference on System Sciences (HICSS'03) - Track1*, page 31.3, Washington, DC, USA. IEEE Computer Society.

Knoth, P. (2008). Information extraction from biomedical texts. Master's thesis, Brno University of Technology.

Knoth, P., Schmidt, M., and Smrz, P. (2008). Information extraction - state of the art.

Knoth, P., Schmidt, M., Smrz, P., and Zdrahal, Z. (2009). Towards a framework for comparing automatic term recognition methods.

Kolak, O. and Schilit, B. N. (2008). Generating links by mining quotations. In *HT '08: Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, pages 117–126, New York, NY, USA. ACM.

Kontopoulos, E., Vrakas, D., Kokkoras, F., Bassiliades, N., and Vlahavas, I. (2008). An ontology-based planning system for e-course generation. *Expert Syst. Appl.*, 35(1-2):398–406.

Kozakov, L., Park, Y., Fin, T., Drissi, Y., Doganata, Y., and Cofino, T. (2004). Glossary extraction and utilization in the information search and delivery system for ibm technical support. *IBM Syst. J.*, 43(3):546–563.

Lakkaraju, P., Gauch, S., and Speretta, M. (2008). Document similarity based on concept tree distance. In *HT '08: Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, pages 127–132, New York, NY, USA. ACM.

Levene, M. and Loizou, G. (2003). Computing the entropy of user navigation in the web. *International Journal of Information Technology and Decision Making*, 2(3):459–476.

LOM (2005). *IEEE P1484.12.3/D8 - Draft Standard for Learning Technology - Extensible Markup Language Schema Definition Language Binding for Learning Object Metadata*.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press.

Medelyan, O. and Witten, I. H. (2006). Thesaurus based automatic keyphrase indexing. In *JCDL '06: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, pages 296–297, New York, NY, USA. ACM.

Nenadié, G., Ananiadou, S., and McNaught, J. (2004). Enhancing automatic term recognition through recognition of variation. In *COLING '04: Proceedings of the 20th international conference on Computational Linguistics*, page 604, Morristown, NJ, USA. Association for Computational Linguistics.

Patry, A. and Langlais, P. (2005). Corpus-based terminology extraction. http://www.iro.umontreal.ca/~felipe/Papers/paper-tke-2005.pdf.

Penas, A., V. F. and Gonzalo, J. (2001). Corpus-based terminology extraction applied to information access. In *Proceedings of Corpus Linguistics 2001*, Lancaster, UK.

Peterson, D. and Levene, M. (2003). Trails record and navigational learning.

Pidgin, O. and Baker, T. (1997). Dublin core in multiple languages.

Porter, M. F. (2000). Java implementation of porter's algorithm.

Prud'hommeaux, E. and Seaborne, A. (2007). Sparql query language for rdf (working draft). Technical report, W3C.

Reynar, J. (1998). *Topic Segmentation: Algorithms and Applications*. PhD thesis.

Rosario, B. (2005). Extraction of semantic relations from bioscience text.

Sclano, F. and Velardi, P. (2007). Termextractor: a web application to learn the shared terminology of emergent web communities. In *Proceedings of the 3rd International Conference on Interoperability for Enterprise Software and Applications (I-ESA 2007)*, Funchal (Madeira Island), Portugal.

SCORM (2009). *Shareable Content Object Reference Model (SCORM) Version 1.0*. Advanced Distributed Learning.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47.

Shipman, III, F. M., Furuta, R., Brenner, D., Chung, C.-C., and Hsieh, H.-w. (1998). Using paths in the classroom: experiences and adaptations. In *HYPERTEXT '98: Proceedings of the ninth ACM conference on Hypertext and hypermedia : links, objects, time and space—structure in hypermedia systems*, pages 267–270, New York, NY, USA. ACM.

Simon, B., Massart, D., Assche, F. V., Ternier, S., Duval, E., Brantner, S., Olmedilla, D., and Mikls, Z. (2005). Z.: A simple query interface for interoperable learning repositories. In *Proceedings of the 1st Workshop on Interoperability of Web-based Educational Systems*, pages 11–18.

Tiun, S., Abdullah, R., and Kong, T. E. (2001). Automatic topic identification using ontology hierarchy. In *CICLing '01: Proceedings of the Second International Conference on Computational Linguistics and Intelligent Text Processing*, pages 444–453, London, UK. Springer-Verlag.

Trigg, R. (1983). *A Network-Based Approach to Text Handling for the Online Scientific Community*. PhD thesis.

Urbán, M.-n. S. and Barriocanal, E. G. (2003). On the integration of ieee-lom metadata instances and ontologies.

Vivaldi, J., Màrquez, L., and Rodríguez, H. (2001). Improving term extraction by system combination using boosting. *Lecture Notes in Computer Science*, 2167:515–521.

Welty, C. and Murdock, J. W. (2006). Towards knowledge acquisition from information extraction. In *5th International Semantic Web Conference (ISWC2006)*.

Wheeldon, R. and Levene, M. (2003). The best trail algorithm for assisted navigation of web sites. In *LA-WEB '03: Proceedings of the First Conference on Latin American Web Congress*, page 166, Washington, DC, USA. IEEE Computer Society.

Wilkinson, R. and Smeaton, A. F. (1999). Automatic link generation. *ACM Computing Surveys*, 31.

Wille, R. (2005). Formal concept analysis as mathematical theory of concepts and concept hierarchies. In [Ganter et al., 2005], pages 1–33.

Yahoo! (2008). Yahoo! content analysis web services: Term extraction.

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *ACL*, pages 189–196.

Zdrahal, Z., Knoth, P., Collins, T., and Mulholland, P. (2009). Reasoning across multilingual learning resources in human genetics. In *Proceedings of ICL 2009*.

Zeng, J. and Bloniarz, P. A. (2004). From keywords to links: an automatic approach. *Information Technology: Coding and Computing, International Conference on*, 1:283.

Zhang, K. and Shasha, D. (1989). Simple fast algorithms for the editing distance between trees and related problems. *SIAM J. Comput.*, 18(6):1245–1262.

Ziqi Zhang, Jose Iria, C. B. and Ciravegna, F. (2008). A comparative evaluation of term recognition algorithms. In (ELRA), E. L. R. A., editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.