

UNIVERSITÀ DI PISA

Scuola di Dottorato in Ingegneria “Leonardo da Vinci”



Corso di Dottorato di Ricerca in INGEGNERIA DELL'INFORMAZIONE

Tesi di Dottorato di Ricerca

Collaborative editing of knowledge resources for cross-lingual text mining

Autore:

Francesco Ronzano _____

Relatori:

Ing. Alessio Bechini _____

Dott. Andrea Marchetti _____

Anno 2011

ABSTRACT

The need to smoothly deal with textual documents expressed in different languages is increasingly becoming a relevant issue in modern text mining environments. Recently the research on this field has been considerably fostered by the necessity for Web users to easily search and browse the growing amount of heterogeneous multilingual contents available on-line as well as by the related spread of the Semantic Web. A common approach to cross-lingual text mining relies on the exploitation of sets of *properly structured multilingual knowledge resources*. The involvement of huge communities of users spread over different locations represents a valuable aid to create, enrich, and refine these knowledge resources. *Collaborative editing Web environments* are usually exploited to this purpose.

This thesis analyzes the features of several knowledge editing tools, both semantic wikis and ontology editors, and discusses the main challenges related to the design and development of this kind of tools. Subsequently, it presents the design, implementation, and evaluation of the Wikyoto Knowledge Editor, called also Wikyoto. Wikyoto is the collaborative editing Web environment that enables Web users lacking any knowledge engineering background to edit the multilingual network of knowledge resources exploited by KYOTO, a cross-lingual text mining system developed in the context of the KYOTO European Project.

To experiment real benefits from social editing of knowledge resources, it is important to provide common Web users with simplified and intuitive interfaces and interaction patterns. Users need to be motivated and properly driven so as to supply information useful for cross-lingual text mining. In addition, the management and coordination of their concurrent editing actions involve relevant technical issues.

In the design of Wikyoto, all these requirements have been considered together with the structure and the set of knowledge resources exploited by KYOTO. Wikyoto aims at enabling common Web users to formalize cross-lingual knowledge by exploiting *simplified language-driven interactions*. At the same time, Wikyoto generates the set of complex knowledge structures needed by computers to mine information from textual contents. The learning curve of Wikyoto has been kept as shallow as possible by hiding the complexity of the knowledge structures to the users. This goal has been pursued by both enhancing the simplicity and interactivity of knowledge editing patterns and by using natural language interviews to carry out the most complex knowledge editing tasks. In this context, TMEKO, a methodology useful to support users to easily formalize cross-lingual information by natural language interviews has been defined. The collaborative creation of knowledge resources has been evaluated in Wikyoto.

INDEX

LIST OF FIGURES AND TABLES	1
I. INTRODUCTION	5
I.I CONTRIBUTION AND SIGNIFICANCE.....	9
I.II OUTLINE	11
1. KNOWLEDGE RESOURCES: BACKGROUND KNOWLEDGE TO SEMANTICALLY STRUCTURE WEB CONTENTS	15
1.1 WEB INFORMATION EXPLOSION: MASSIVE, HETEROGENEOUS, USER GENERATED, MULTILINGUAL CONTENTS	17
1.2 NATURAL LANGUAGE PROCESSING AND THE SEMANTIC WEB: MINING, STRUCTURING AND INTEGRATING CONTENTS ON A WEB SCALE BY MEANS OF DATA SEMANTICS	20
1.2.1 Mining textual contents through Natural Language Processing.....	21
1.2.2 Semantically describing and interlinking data on a Web scale: the Semantic Web.....	26
1.2.2.1 Data formats for semantic descriptions of on-line contents: RDF and OWL.....	28
1.2.2.2 The URI system: unambiguously refer and retrieve (semantic) descriptions of informative and non-informative resources all over the Web	30
1.2.2.3 Semantic Web Search Engines: searching for semantic data over the Web.....	35
1.2.2.4 Linked Data: a Web of interlinked distributed semantic datasets .	36
1.2.3 Natural Language Processing underpins the Semantic Web	37
1.3 KNOWLEDGE RESOURCES: BACKGROUND KNOWLEDGE TO SUPPORT WEB DATA SEMANTICS	38
1.3.1 Taxonomy of knowledge resources	39
1.3.2 Computational lexicons: mining semantics from texts.....	41
1.3.2.1 WordNet.....	43
<i>WordNet and the management of multilingual contents</i>	48
1.3.3 Ontologies: representing and reasoning about data	49
1.3.3.1 OWL: the DL formalism to define Web Ontologies.....	51
<i>Reasoning procedures based on DL.....</i>	53
<i>The Web Ontology Language</i>	54
2. EDITING KNOWLEDGE RESOURCES: THE WIKI WAY	61
2.1 THE WIKI PARADIGM APPLIED TO KNOWLEDGE RESOURCES..	62
2.2 ENVIRONMENTS TO EDIT KNOWLEDGE RESOURCES	64
2.2.1 Wiki editors of textual contents enriched with semantic annotations..	65

Platypus Wiki	65
Semantic MediaWiki.....	66
IKEWiki and the KiWi project.....	67
OntoWiki	69
Rizhome.....	70
SweetWiki.....	70
Maariwa.....	71
SAVVY Wiki	72
2.2.2 Ontology editors.....	72
2.2.2.1 Ontology editors based on a graphical interface	72
<i>Collaborative Protégé and Web Protégé</i>	<i>73</i>
<i>Ontostudio</i>	<i>74</i>
<i>The NeOn Toolkit and the NeOn Project.....</i>	<i>75</i>
<i>Ontoverse</i>	<i>76</i>
<i>TopBraid Composer</i>	<i>77</i>
<i>CODA.....</i>	<i>78</i>
<i>SWOOP</i>	<i>79</i>
2.2.2.2 Ontology editors and ontology editing methodologies based on controlled languages	79
<i>The Attempto Controlled Language and ACEwiki.....</i>	<i>80</i>
<i>CLOnE and the RoundTrip Ontology Authoring</i>	<i>81</i>
<i>GINO</i>	<i>82</i>
2.3 COMPARING KNOWLEDGE EDITORS	83
2.3.1 Analysis of semantic wikis	83
2.3.2 Analysis of ontology editors	84
2.3.3 The desirable features of a collaborative knowledge editor.....	86
2.4 USER MOTIVATION IN COLLABORATIVE KNOWLEDGE EDITING	88
3. WIKYOTO KNOWLEDGE EDITOR: THE COLLABORATIVE WEB ENVIRONMENT TO MANAGE KYOTO KNOWLEDGE RESOURCES	95
3.1 KYOTO: A CROSS-LINGUAL TEXT MINING ENVIRONMENT	97
3.1.1 Knowledge based cross-lingual text mining in KYOTO	97
3.1.2 The knowledge resources of KYOTO: the Multilingual Knowledge Base	100
3.1.2.1 The data formats of KYOTO knowledge resources: OWL and WordNet-LMF.....	102
3.1.2.2 Mapping relations among KYOTO knowledge resources	103
3.1.3 The architecture of KYOTO.....	105
3.2 WIKYOTO KNOWLEDGE EDITOR	109
3.2.1 Collaborative editing of the Multilingual Knowledge Base: motivations	109
3.2.2 Shaping Wikyoto: system design issues	112
3.2.2.1 Editing actions targeted to gather linguistic information useful to support cross-lingual text mining	112

3.2.2.2 Exploitation of external knowledge resources to enrich the Multilingual Knowledge Base: the KYOTO Terminology, SKOS Thesauri, and DBpedia.....	115
3.2.2.3 Intuitive visualization and simplified language-driven user interaction patterns to browse and edit knowledge resources.....	117
3.2.2.4 Need for a collaborative tool balancing between semantic wikis and ontology editors.....	120
3.2.3 Wikyoto architecture and implementation	121
3.2.3.1 Software design concerns of browser-based real-time collaborative editing systems	122
3.2.3.2 The architecture of Wikyoto	123
3.2.3.3 Data Repositories	126
<i>KYOTO Data Repositories</i>	126
<i>External Data Repositories</i>	127
3.2.3.4 Browser Module: Wikyoto interface and Javascript libraries	129
<i>Wikyoto interface layout</i>	129
<i>Javascript client-side elaborations</i>	132
3.2.3.5 Managing concurrent editing actions in Wikyoto	134
3.2.3.6 The implementation technologies of Wikyoto.....	136
3.2.4 Exploiting Wikyoto	137
3.2.5 TMEKO: supporting users to formalize cross-lingual information	140
3.2.5.1 Mapping WordNet synsets to the KYOTO Central Ontology.....	141
3.2.5.2 The steps of the TMEKO procedure	145
3.2.5.3 TMEKO and TMEO: language-driven vs. logic-driven approaches to enrich ontologies.....	148
3.2.6 Evaluation	149
4. CONCLUSIONS AND PERSPECTIVES	155
BIBLIOGRAPHY	159
APPENDIX: LIST OF PUBLICATIONS	171

