Semantics in Adaptive and Personalised Systems

Pasquale Lops · Cataldo Musto ·
Fedelucio Narducci · Giovanni Semeraro

Semantics in Adaptive and Personalised Systems

Methods, Tools and Applications

Springer

Pasquale Lops
Dipartimento di Informatica
Università di Bari Aldo Moro
Bari, Italy

Cataldo Musto
Dipartimento di Informatica
Università di Bari Aldo Moro
Bari, Italy

Fedelucio Narducci
Dipartimento di Informatica
Università di Bari Aldo Moro
Bari, Italy

Giovanni Semeraro
Dipartimento di Informatica
Università di Bari Aldo Moro
Bari, Italy

ISBN 978-3-030-05617-9 ISBN 978-3-030-05618-6 (eBook)
https://doi.org/10.1007/978-3-030-05618-6

© Springer Nature Switzerland AG 2019
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

To my kids Giuseppe and Annapaola,
love of my life.
Pasquale Lops

To those who taught me about the importance
of the "meaning". In every thing.
Cataldo Musto

To women and men who change our lives
for the better.
Fedelucio Narducci

In memoriam Aaron Hillel Swartz and Enzo
Roberto Tangari.
Giovanni Semeraro

Foreword

Web search engines and recommender systems are among today's most visible and successful applications of Artificial Intelligence technology in practice. We rely on such systems every day when we search for information, when we shop online, or when we stream videos. Without such systems, it almost seems impossible to find things that interest us within the huge amounts of information that are offered online today.

Research in the area of Information Filtering and Information Retrieval dates back to the 1970s or even earlier. One central task of such systems then and today is to estimate to what extent a given document or web page is relevant for a given query by the user. Although the final ranking of the relevant documents is often influenced by other factors, e.g., the general popularity of a web page, any search system at some stage makes inferences about what each indexed web page is about, i.e., about its content.

Over the decades, these forms of reasoning have become more and more sophisticated. On one hand, different internal document representations were developed, from simple term-counting approaches, over latent semantic approaches to embedding models, which implicitly encode semantic relationships between terms. At the same time, more and more structured or unstructured external knowledge sources have become available, e.g., in the form of Linked Data, which allow search and information filtering systems to make inferences using explicitly given semantic relations between the concepts that appear in queries and in documents.

The same set of techniques can also be applied in the field of recommender systems, which is the main focus of this book. Here, the input to ranking task is not an individual query, but a user profile that the system has learned from past user interactions over time. Accordingly, such content-based or semantics-based recommenders are able to personalize the ranking on the bases of the assumed user interests.

In the traditional categorization of recommendation techniques, content-based methods (here, the term content also covers metadata and other side information) are often considered as an alternative to collaborative filtering approaches. This latter class of systems, which base their recommendations on behavioral patterns of

a larger user community, dominates the research landscape. However, pure collaborative approaches can have a number of limitations. It is, for example, difficult to ensure that a set of recommendations is diverse when we do not know anything about the similarity of two items. Likewise, explaining recommendations to users can be a challenge when we cannot inform users how the attributes of a recommended item relate to their preferences. In a number of application domains, it is therefore favorable to design a hybrid system that combines knowledge about the items with collaborative information.

The literature on semantics-based or hybrid recommender systems is actually quite rich, but unfortunately also scattered. Today, relevant works appear in publication outlets of different communities, e.g., Information Retrieval, Semantic Web, or Recommender Systems. This book therefore fills an existing gap in the literature. It first provides an introduction to the basic concepts of content representations, then discusses approaches for semantic analysis, and covers today's external knowledge sources that can be leveraged for information filtering and recommendation. Based on these foundations, it then reviews use cases of how rich content information can be used to build better recommenders for the future.

Klagenfurt, Austria Dietmar Jannach
June 2019

Preface

The human desire to make the machine always smarter has been the driving force for all the research in the Artificial Intelligence (AI) area.

Generally speaking, what makes a system *intelligent* is the capability of *understanding* signals coming from the environment and of correctly adapting its behavior accordingly. Such a capability is strictly related to the definition and the design of specific techniques for *interpreting* messages generated by the users.

Some years ago, when we typed on Google the query *How tall is the Eiffel Tower?*, the system answered with a set of documents, some of them including the information we were seeking for, but without a precise identification of the correct answer. Today, this is no longer the case since intelligent assistants like Siri, Alexa, or Google assistant, and the Google search engine itself, are able to provide the *exact answer* the user is looking for, that is, in the case of the Eiffel Tower, *300 m*.

Without any doubt, we can state that the *semantics* represents the theoretical foundation to implement models and technologies that allow the machines to interpret and understand information provided in natural language. Indeed, thanks to the semantics, it is possible to give meaning to documents, sentences, and questions expressed in natural language and to create a bridge between the *information needs* of a user and the *answers* to those needs.

Such an intuition is currently implemented in several tools and platforms as search engines, recommender systems, digital assistants, and contributes to their tangible improvement in accuracy and effectiveness we are recently witnessing.

We hope this book could become a reference point in the panorama of *adaptive and personalized systems* exploiting semantics. The book is organized into three main parts. First, we motivate the need to exploit textual content in intelligent information access systems, and then we give an overview of the basic methodologies to process and represent content-based features. Next, we thoroughly describe state-of-the-art methodologies and techniques to enrich textual content representation by introducing semantics. Finally, the last part of the book provides a more *practical perspective* and discusses several applications that exploit the techniques introduced and described in the previous chapters.

We would like to sincerely thank everyone who contributed to this book, and the various people who provided us with comments and suggestions and encouraged us to summarize years of work in a single book. We thank, in particular, Nancy Wade-Jones from Springer, who supported us throughout the editorial process.

We are very grateful to the people of the Semantic Web Access and Personalization—SWAP research group,[1] who contributed to most of the work cited and described in this book. We would like to thank Marco de Gemmis, who started to investigate how Natural Language Processing techniques could be adopted to devise a new generation of content-based recommender systems, Pierpaolo Basile, who made available his great expertise related to Word Sense Disambiguation and Distributional Semantics Models, which were successfully used in complex recommendation environments, Annalina Caputo, a former member of the research group working on semantic information retrieval methods. We would also like to thank all the other collaborators, Ph.D. students, and research fellows of the SWAP research group, in particular, Leo Iaquinta, Andrea Iovine, Piero Molino, Marco Polignano, Gaetano Rossiello, Lucia Siciliani, and Vincenzo Tamburrano, each giving a specific contribution to the ideas, systems, and research presented in this book.

Bari, Italy
July 2019

Pasquale Lops
Cataldo Musto
Fedelucio Narducci
Giovanni Semeraro

[1] http://www.di.uniba.it/%7Eswap/.

Contents

1 Introduction 1
1.1 Data Explosion and Information Overload 2
1.2 Intelligent Information Access 5
1.2.1 Information Retrieval and Information Filtering 5
1.2.2 Recommender Systems 7
1.3 Why Do We Need Content? 11
1.3.1 Tackling the Issues of Collaborative Filtering 11
1.3.2 Feed and Follow Recent Trends 15
1.4 Why Do We Need Semantics? 16
References 19

2 Basics of Content Representation 21
2.1 Pipeline for Text Processing 22
2.1.1 Lexical Analysis 23
2.1.2 Syntactic Analysis 31
2.2 Vector Space Model 35
2.3 Semantics-Aware Content Representation 40
References 42

3 Encoding Endogenous Semantics 43
3.1 Distributional Semantics Models 44
3.2 Word Embedding Techniques 49
3.2.1 Latent Semantic Analysis 50
3.2.2 Random Indexing 53
3.2.3 Word2Vec 57
3.3 Explicit Semantic Analysis 61
References 66

4 Encoding Exogenous Semantics 71
4.1 Overview of Structured Knowledge Sources 72
4.1.1 WordNet 73
4.1.2 BabelNet: An Encyclopedic Dictionary 75
4.1.3 Linked Open Data and DBpedia 78
4.1.4 Wikidata 82
4.2 Linking Item Features to Concepts 84
4.2.1 Word Sense Disambiguation 85
4.2.2 Entity Linking 91
4.3 Linking Items to a Knowledge Graph 94
4.3.1 Use of Ontologies 95
4.3.2 Use of Linked Open Data 98
References 102

5 Adaptive and Personalized Systems Based on Semantics 105
5.1 Semantics-Aware Recommender Systems 105
5.1.1 Approaches Based on Endogenous Semantics 106
5.1.2 Approaches Based on Exogenous Semantics 112
5.1.3 Semantics-Aware User Profiling Techniques 125
5.2 Semantics-Aware Social Media Analysis 129
5.2.1 L'Aquila Social Urban Network 132
5.2.2 The Italian Hate Map 136
5.3 New Trends and Challenges 138
5.3.1 Cross-Lingual Recommendations 138
5.3.2 Conversational Recommender Systems 144
5.3.3 Explaining Recommendations 149
5.3.4 Serendipitous Recommendations 158
References 161

6 Conclusions and Future Challenges 169
6.1 Conclusions 169
6.2 Future Challenges 170
References 171

Appendix: Available Tools and Resources 173

Acronyms

AI	Artificial Intelligence
BNC	British National Corpus
BOW	Bag of Words
CBOW	Continuous Bag of Words
CBRS	Content-Based Recommender System
CF	Collaborative Filtering
CL-ESA	Cross-Language Explicit Semantic Analysis
CoRS	Conversational Recommender System
DM	Dialog Manager
DSM	Distributional Semantics Model
EL	Entity Linking
EPG	Electronic Program Guides
ER	Entity Recognition
ESA	Explicit Semantic Analysis
GDPR	General Data Protection Regulation
IDF	Inverse Document Frequency
IF	Information Filtering
IMDB	Internet Movie Database
IR	Information Retrieval
KB	Knowledge Base
LDA	Latent Dirichlet Allocation
LOD	Linked Open Data
LSA	Latent Semantic Analysis
LSI	Latent Semantic Indexing
MSS	Most Specific Subsumer
MUC	Message Understanding Conference
NER	Named Entity Recognition
NLP	Natural Language Processing
NMF	Nonnegative Matrix Factorization
NNDB	Notable Names Database

NP	Noun Phrases
OWL	Ontology Web Language
PCA	Principal Component Analysis
pLSA	Probabilistic Latent Semantic Analysis
PMI	Pointwise Mutual Information
POS	Part of Speech
QA	Question Answering
RDF	Resource Description Framework
RI	Random Indexing
RP	Random Projection
RS	Recommender Systems
SA	Sentiment Analyzer
SG	Skip-Gram
SPARQL	SPARQL Protocol and RDF Query Language
SUN	Social Urban Network
SVD	Singular Value Decomposition
TF	Term Frequency
TR-ESA	Translation-based Explicit Semantic Analysis
URI	Uniform Resource Identifier
VP	Verb Phrases
VSM	Vector Space Model
WSD	Word Sense Disambiguation

List of Figures

Fig. 1.1 Workflow carried out by a generic search engine 6
Fig. 1.2 Workflow carried out by a generic information filtering tool 6
Fig. 1.3 Workflow carried out by a recommender system 8
Fig. 1.4 Toy example of a data model for a collaborative recommender system, based on the user–item matrix 12
Fig. 1.5 Issues of collaborative recommender systems: sparsity and new item problem 13
Fig. 1.6 A content-based recommendation pipeline 14
Fig. 1.7 Ambiguity in user modeling and recommendations 16
Fig. 1.8 Limits of keyword-based representation in recommendation tasks 17
Fig. 1.9 Vocabulary mismatch problem 18
Fig. 2.1 The natural language processing pipeline 23
Fig. 2.2 Named entities recognized in the text fragment 29
Fig. 2.3 Penn treebank tagset 30
Fig. 2.4 POS tagging for a text fragment 31
Fig. 2.5 Chunking for a text fragment. NP = noun phrase, VP = verb phrase 32
Fig. 2.6 A parse tree for a text fragment 34
Fig. 2.7 The term–document matrix reporting the number of times each word (row) occurs in each document (column). 36
Fig. 2.8 The term–document matrix reporting the TF-IDF weight for each word (row) in each document (column) 38
Fig. 2.9 A graphical representation of cosine similarity 39
Fig. 2.10 Items and user profiles represented in a vector space 40
Fig. 2.11 Classification of semantic representation techniques 41
Fig. 3.1 Similar terms share similar usages 45
Fig. 3.2 The term–context matrix 46
Fig. 3.3 A two-dimensional representation of a `WordSpace` 47
Fig. 3.4 An example of term–sentence matrix 48

Fig. 3.5 An example of term–term matrix 48
Fig. 3.6 An example of SVD applied to the term–document matrix 52
Fig. 3.7 A visual explanation of the Johnson–Lindenstrauss lemma. *Z* is the nearest point to *X* in the reduced vector space, as in the original space, even though the *numerical* value of their pairwise similarity is different 54
Fig. 3.8 Context vectors of dimension k = 8. 56
Fig. 3.9 The vector space representation of a term obtained by summing the context k-dimensional index vectors the term co-occurs with 56
Fig. 3.10 Uniform representation of `WordSpace` and `DocSpace` 57
Fig. 3.11 Structure of the network 59
Fig. 3.12 Continuous Bag-of-Words methodology 59
Fig. 3.13 Skip-Gram methodology 60
Fig. 3.14 The ESA matrix 62
Fig. 3.15 Semantics of the word *cat* 63
Fig. 3.16 Attribute vector of the *Panthera* Wikipedia article 63
Fig. 3.17 Semantic relatedness between semantic interpretation vectors 64
Fig. 3.18 An example of semantic interpretation vector of a text fragment 65
Fig. 3.19 The semantic interpretation vector of *The Matrix* 65
Fig. 4.1 The hierarchy of sense 1 of the word "bat" obtained from WordNet (version 2.1) 74
Fig. 4.2 Apple in the sense of fruit in BabelNet. 76
Fig. 4.3 Apple in the sense of multinational corporation in BabelNet. 77
Fig. 4.4 The Linked Open Data cloud. Each bubble represents a dataset (a set of `RDF` statements). Datasets encoding similar or related information are represented with the same colors. 79
Fig. 4.5 The so-called Semantic Web Cake. Each element in the cake represents a formalism or a technology that was necessary to enable the vision of the Semantic Web. 80
Fig. 4.6 An example of `RDF` triple, encoding the information that Keanu Reeves has acted in The Matrix. The `URI` `dbr:Keanu_Reeves` is an abbreviation for `http://dbpedia.org/resource/resource/Keanu_Reeves` 80
Fig. 4.7 Data Mapping between Wikipedia and DBpedia. 81
Fig. 4.8 A (tiny) portion of the properties, available in the LOD cloud, that describe the band *The Coldplay*. 82
Fig. 4.9 A portion of the data available in Wikidata that describe the band *The Coldplay* 83

Fig. 4.10 The preprocessing of sentence "The white cat is hunting the mouse." Each token is labeled with a tag describing its lexical role in the sentence. NN = noun, singular—VB = verb, base form—JJ = adjective . . . 87
Fig. 4.11 A fragment of the WordNet taxonomy . . . 88
Fig. 4.12 Similarities between synsets . . . 88
Fig. 4.13 The synset–document matrix reporting the number of times each synset (row) occurs in each document (column) . . . 91
Fig. 4.14 An example of entity linking performed by Tagme . . . 93
Fig. 4.15 An example of entity linking performed by Babelfy . . . 93
Fig. 4.16 An example of entity linking performed by DBpedia Spotlight . . . 94
Fig. 4.17 A tiny portion of the *Movie Ontology* . . . 96
Fig. 4.18 A tiny portion of the *Movie Ontology* with instances . . . 96
Fig. 4.19 Genre class hierarchy in Movie Ontology . . . 97
Fig. 4.20 An excerpt of the Quickstep research paper topic ontology . . . 98
Fig. 4.21 An example of SPARQL query . . . 99
Fig. 4.22 Mapping between items and DBpedia URIs . . . 100
Fig. 4.23 A data model including the features extracted from the Linked Open Data cloud . . . 101
Fig. 5.1 An example of enrichment by ESA . . . 111
Fig. 5.2 An example of a keyword-based profile . . . 114
Fig. 5.3 An example of a synset-based profile . . . 114
Fig. 5.4 Neighborhood formation from clustered partitions . . . 115
Fig. 5.5 Sample RDF graph extracted from DBpedia and LinkedMDB . . . 117
Fig. 5.6 Matrix representation of property resource indexes . . . 117
Fig. 5.7 Vector space model for LOD . . . 118
Fig. 5.8 Basic bipartite graph representing users, items, and their preferences . . . 118
Fig. 5.9 Tripartite graph representing users, items, and information coming from the LOD cloud . . . 120
Fig. 5.10 Examples of user profiles . . . 127
Fig. 5.11 Workflow for Holistic user modeling carried out by Myrror . . . 128
Fig. 5.12 Semantic Tag cloud showing users' interests . . . 129
Fig. 5.13 The architecture of the semantic content analysis framework . . . 131
Fig. 5.14 Example of data visualizations available in CrowdPulse . . . 133
Fig. 5.15 Social capital indicators . . . 133
Fig. 5.16 Example of ambiguous Tweets . . . 134
Fig. 5.17 Entity-based representation of a Tweet . . . 134
Fig. 5.18 An example of the workflow carried out by the content scoring and classification module of L'Aquila social urban network project . . . 135

Fig. 5.19 An example of the output returned by the pipeline of algorithms implemented in the Italian hate map project 136
Fig. 5.20 Alignment of synsets in different languages in MultiWordNet 140
Fig. 5.21 Example of synset-based document representation in different languages 140
Fig. 5.22 Multilingual document representation using translation-based ESA (TR-ESA). 141
Fig. 5.23 Multilingual document representation using cross-language ESA (CL-ESA). 142
Fig. 5.24 Wikipedia cross-language links (in the red box). 143
Fig. 5.25 The general architecture of a conversational recommender system. 147
Fig. 5.26 The Bot workflow. 149
Fig. 5.27 A screenshot of the Bot during the training phase in typing mode (**a**) and the recommendation phase (**b**). 150
Fig. 5.28 The histogram with grouping interface, which performed best in the study of Herlocker et al. [46] 152
Fig. 5.29 Personalized tag cloud as defined in [37] 152
Fig. 5.30 Workflow carried out by our framework. 154
Fig. 5.31 Toy example related to the modeling of *direct* connections 157
Fig. 5.32 Toy example related to modeling of *indirect* connections. 158
Fig. 5.33 Fragment of the row of the correlation matrix for the movie Star Wars. Each cell reports the correlation index between Star Wars and the movie on the column, and the set of plot keywords which match the new keywords produced by the KI process 161