Semantic
Web
Access and
Personalization
research group
http://www.di.uniba.it/~swap

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

Semantics-aware Content-based Recommender Systems

the Big Picture

Pasquale Lops

Dept. of Computer Science
University of Bari "Aldo Moro", Italy

1st Workshop on NEW TRENDS IN CONTENT-BASED RECOMMENDER SYSTEMS
CBRecSys 2014 - Foster City, CA, US – October 6, 2014

DEC 13

On the uselessness of content for recommendations

This is one of the hot discussions that has sparked as a result of the Netflix Prize. During the competition several teams reported trying to use movie metadata always with discouraging results. This is probably best summarized by a 2008 post by Pragmatic Theory, one of the leading teams.

The issue was re-opened during the last Recsys conference in two ways. First, there was an interesting discussion during one of the panels including the leading teams. Second a paper with a rather provocative title was published: "Recommending new movies: even a few ratings are more valuable than metadata".

After this, I have seen several discussions in which people use these findings to conclude that content-based recommendations are little more than a dead end, and it is not worth investing in such research. One such discussion happened in the Recommender Systems group in LinkedIn, but it was in the Music-IR list, where things heated up the most, turning into a long and interesting thread. Most of what follows is basically an edited version of what I already expressed in those two discussions.

BAD NEWS

Recommending New Movies: Even a Few Ratings Are More Valuable Than Metadata

István Pilászy
Dept. of Measurement and Information Systems
Budapest University of Technology and Economics
Magyar Tudósok krt. 2.
Budapest, Hungary
pila@mit.bme.hu

Domonkos Tikk
Dept. of Telecom. and Media Informatics
Budapest University of Technology and Economics
Magyar Tudósok krt. 2.
Budapest, Hungary
tikk@tmit.bme.hu

ABSTRACT
The Netflix Prize (NP) competition gave much attention to collaborative filtering (CF) approaches. Matrix factorization (MF) based CF approaches assign low dimensional feature vectors to users and items. We link CF and content-based filtering (CBF) by finding a linear transformation that

1. INTRODUCTION
The goal of recommender systems is to give personalized recommendation on items to users. Typically the recommendation is based on the former and current activity of the users, and metadata about users and items, if available.
There are two basic strategies that can be applied when

Workshop on New Trends in Content-based Recommender Systems
CBRecSys 2014 | RecSys 2014, Silicon Valley, US | October 6, 2014

GOOD NEWS

Organization

Organizers

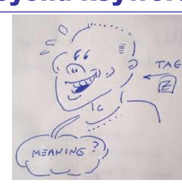
- Toine Bogers (toine@hum.aau.dk), Aalborg University Copenhagen, Denmark
- Marijn Koolen (marijn.koolen@uva.nl), University of Amsterdam, the Netherlands
- Iván Cantador (ivan.cantador@uam.es), Universidad Autónoma de Madrid, Spain

---

Beyond keywords

Semantic Text Analytics

Semantic	Analytics
concept identification in text-based representations through advanced NLP techniques **"beyond keywords"**	Machine Learning & Data Mining–based Personalization **"deep user profiles"**



Part I: **Content-based Recommender Systems (CBRS): basics**

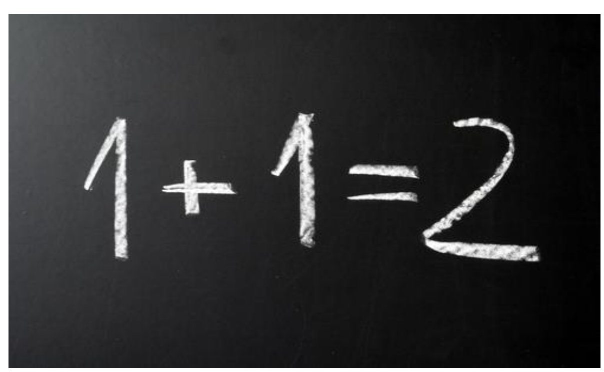
Part II: **Limited content analysis**

Beyond keywords: **semantics** into CBRS

Part III: **Overspecialization**

Serendipitous recommendations

Outline



Basics

Content-based RecSys (CBRS)

Suggest items similar to those the user liked in the past

Recommendations generated by matching the **description of items** with the **profile of the user's interests**

use of specific **features**

[Lops11] P. Lops, M. de Gemmis, and G. Semeraro. Content-based recommender systems: State of the art and trends. In F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor (Eds.), *Recommender Systems Handbook*, Springer, 73-105, 2011.

[Pazzani07] Pazzani, M. J., & Billsus, D. Content-Based Recommendation Systems. *The Adaptive Web.* Lecture Notes in Computer Science vol. 4321, 325-341, 2007.

Advantages

USER INDEPENDENCE

No need of ratings from the community to bootstrap CBRS

TRANSPARENCY

Simple explanations by listing content-features that caused an item to be recommended

Drawback

LIMITED CONTENT ANALYSIS

no suitable suggestions if no enough information in the content to **discriminate** items the user **likes** / **does not like**

need of **meaningful features**

keywords not appropriate for representing content
multi-word concepts, **synonymy**, **polysemy**

Keyword-based Profiles

doc1: **AI** is a branch of computer science

doc2: the 2011 International Joint Conference on **Artificial Intelligence** will be held in Spain

doc3: **apple** launches a new product…

USER PROFILE

artificial	0.02
intelligence	0.01
apple	0.13
AI	0.15
…	

MULTI-WORD CONCEPTS

Keyword-based Profiles

doc1: **AI** is a branch of computer science

doc2: the 2011 International Joint Conference on **Artificial Intelligence** will be held in Spain

doc3: **apple** launches a new product…

USER PROFILE

artificial	0.02
intelligence	0.01
apple	0.13
AI	0.15
…	

SYNONYMY

Keyword-based Profiles

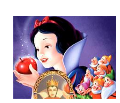
doc1: **AI** is a branch of computer science

doc2: the 2011 International Joint Conference on **Artificial Intelligence** will be held in Spain

doc3: **apple** launches a new product…

USER PROFILE

artificial	0.02
intelligence	0.01
apple	0.13
AI	0.15
…	



POLYSEMY

Advanced methods are needed for the elicitation of user interests

Drawback

OVERSPECIALIZATION

user is going to be recommended
items similar to those already rated (OBVIOUSNESS)

no inherent method for finding something **unexpected**

SERENDIPITY PROBLEM

[McNee06] S.M. McNee, J. Riedl, and J. Konstan. Accurate is not always good: How accuracy metrics have hurt recommender systems. In *Extended Abstracts of the 2006 ACM Conference on Human Factors in Computing Systems*, pages 1–5, Canada, 2006.

Harry Potter for ever?

Part I: **Content-based Recommender Systems (CBRS): basics**

Part II: **Limited content analysis**

Beyond keywords: **semantics** into CBRS

Part III: **Overspecialization**

Serendipitous recommendations

Outline

From CBRSs to semantics-aware CBRSs through Semantic Text Analytics

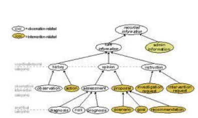
Semantic techniques

Top-down	**Bottom-up**
integration of **external knowledge** for representing items and user profiles providing the recommender with the **cultural background** and **linguistic knowledge**	**meaning of a word** determined by the **rules of its usage** in the context of ordinary and concrete language behaviour
EXPLICIT SEMANTICS	**IMPLICIT SEMANTICS**

Top-down approaches

use of

ontological knowledge from simple **linguistic ontologies** to more complex domain-specific ones	**unstructured encyclopedic knowledge** sources, such as **Wikipedia**	**Linked Open Data cloud**



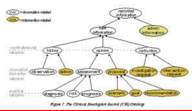
WIKIPEDIA

DBpedia

Top-down approaches

use of

ontological knowledge from simple **linguistic ontologies** to more complex domain-specific ones



unstructured encyclopedic knowledge sources, such as **Wikipedia**

Linked Open Data cloud

From words to meanings: Word Sense Disambiguation (WSD)

WSD selects the proper meaning (**sense**) for a word in a text by taking into account the context in which it occurs

context

Apple $\underbrace{\text{Computer iPhone}}$

#12567: computer brand

~~#22999: fruit~~

Sense Repository — Dictionaries, Ontologies, e.g. WordNet

P. Basile, M. Degemmis, A. Gentile, P. Lops, and G. Semeraro. UNIBA: JIGSAW algorithm for Word Sense Disambiguation. In *Proceedings of the 4th ACL 2007 International Workshop on Semantic Evaluations (SemEval-2007), Prague, Czech Republic*, pages 398–401, Association for Computational Linguistics, June 23–24, 2007.

Sense-based Profiles

doc1: **AI** is a branch of computer science

doc2: the 2011 International Joint Conference on **Artificial Intelligence** will be held in Spain

doc3: **apple** launches a new product…

USER PROFILE

#12387	0.03
apple	0.13
AI	0.15
…	

MULTI-WORD CONCEPTS

Sense-based Profiles

doc1: **AI** is a branch of computer science

doc2: the 2011 International Joint Conference on **Artificial Intelligence** will be held in Spain

doc3: **apple** launches a new product…

USER PROFILE

#12387	0.18
apple	0.13
#12387	0.15
…	

SYNONYMY

Sense-based Profiles

doc1: **AI** is a branch of computer science

doc2: the 2011 International Joint Conference on **Artificial Intelligence** will be held in Spain

doc3: **apple** launches a new product…

SEMANTIC USER PROFILE
sense identifiers rather than keywords

USER PROFILE

#12387	0.18
#12567	0.13
…	

POLYSEMY

M. Degemmis, P. Lops, and G. Semeraro. A Content-collaborative Recommender that Exploits WordNet-based User Profiles for Neighborhood Formation. *User Modeling and User-Adapted Interaction: The Journal of Personalization Research (UMUAI)*, 17(3):217–255, Springer Science + Business Media B.V., 2007.

G. Semeraro, M. Degemmis, P. Lops, and P. Basile. Combining Learning and Word Sense Disambiguation for Intelligent User Profiling. In M. M. Veloso, editor, *IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, January 6-12, 2007*, pages 2856–2861. Morgan Kaufmann, 2007.

M.Degemmis, P. Lops, G. Semeraro, Pierpaolo Basile: Integrating tags in a semantic content-based recommender ACM Conference on Recommender Systems, RecSys 2008: 163-170

Leverage crowdsourcing for deep text analytics

WordNet

- ✓ few named entities
- ✓ no events
- ✓ evolving vocabularies

Wikipedia

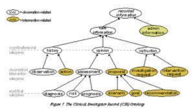
- ✓ free and covers many domains
- ✓ under constant development
- ✓ highly accurate [Giles05]

[Giles05] J. Giles. Internet Encyclopaedias Go Head to Head. *Nature*, 438:900–901, 2005.

Top-down approaches

use of

ontological knowledge from simple **linguistic ontologies** to more complex domain-specific ones	**unstructured encyclopedic knowledge** sources, such as **Wikipedia**	**Linked Open Data cloud**



Explicit Semantic Analysis (ESA)

fine-grained **semantic representation** of natural language texts in a high-dimensional space of **comprehensible concepts** derived from Wikipedia [Gabri06]

Wikipedia viewed as an **ontology**

[Gabri06] E. Gabrilovich and S. Markovitch. Overcoming the Brittleness Bottleneck using Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge. In *Proceedings of the 21th National Conf. on Artificial Intelligence and the 18th Innovative Applications of Artificial Intelligence Conference*, pages 1301–1306. AAAI Press, 2006.

Explicit Semantic Analysis (ESA)

ESA matrix

semantic relatedness between a word and a concept **TFIDF score**

Wikipedia articles

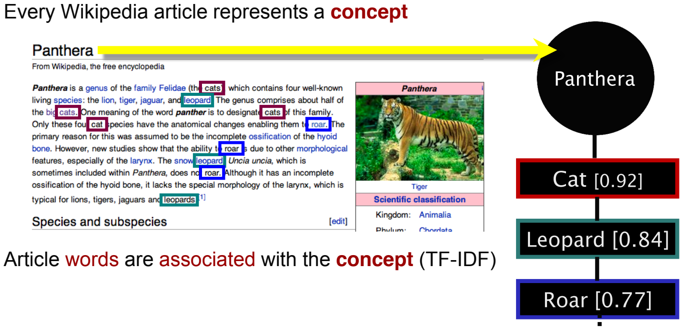
Terms

ESA	Concept 1	...	Concept n
Term 1	TFIDF	TFIDF	TFIDF
...	TFIDF	TFIDF	TFIDF
Term k	TFIDF	TFIDF	TFIDF

Explicit Semantic Analysis (ESA)

Wikipedia is viewed as an **ontology** - a collection of concepts

Every Wikipedia article represents a **concept**

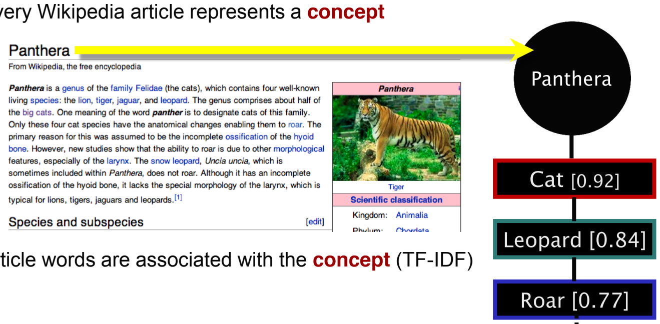


Article words are associated with the **concept** (TF-IDF)

Explicit Semantic Analysis (ESA)

Wikipedia is viewed as an **ontology** - a collection of concepts

Every Wikipedia article represents a **concept**



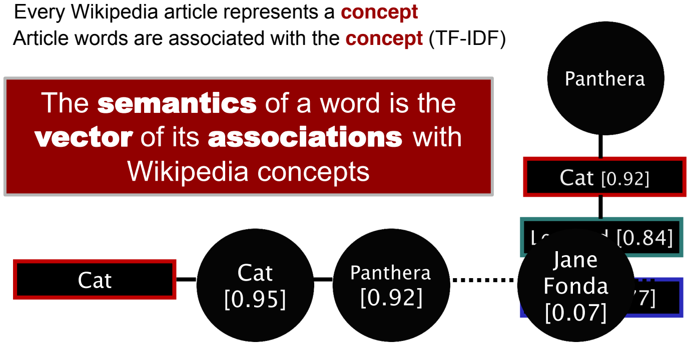
Article words are associated with the **concept** (TF-IDF)

Explicit Semantic Analysis (ESA)

Wikipedia is viewed as an **ontology** - a collection of concepts

Every Wikipedia article represents a **concept**
Article words are associated with the **concept** (TF-IDF)

The **semantics** of a word is the **vector** of its **associations** with Wikipedia concepts



Explicit Semantic Analysis (ESA)

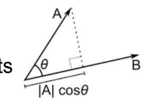
The semantics of a text fragment is the centroid of the semantics of its words

mouse	Mouse (rodent) [0.91]	Mouse (computing) [0.84]	Mickey Mouse [0.81]	...	John Steinbeck [0.17] ...
button	Button [0.93]	Dick Button [0.84]	Mouse (computing) [0.81]	...	Game Controller [0.32] ...
mouse button	Mouse (computing) [0.95]	Drag-and-drop [0.91]	Game Controller [0.64]	...	Mouse (rodent) [0.56] ...

ESA effectively used for

Text Categorization [Gabri09]
experiments on diverse datasets

Semantic relatedness of words and texts [Gabri09]
cosine similarity between vectors of ESA concepts



Information Retrieval [Egozi08, Egozi11]
ESA-based IR algorithm enriching documents and queries

What about ESA for Information Filtering?

[Gabri09] E. Gabrilovich and S. Markovitch. Wikipedia-based Semantic Interpretation for Natural Language Processing. *Journal of Artificial Intelligence Research* 34:443-498, 2009.

[Egozi08] Ofer Egozi, Evgeniy Gabrilovich, Shaul Markovitch: Concept-Based Feature Generation and Selection for Information Retrieval. *AAAI 2008*, 1132-1137, 2008.

[Egozi11] Ofer Egozi, Shaul Markovitch, Evgeniy Gabrilovich. Concept-Based Information Retrieval using Explicit Semantic Analysis. *ACM Transactions on Information Systems* 29(2), April 2011.

Information Filtering using ESA

TV-domain::German Electronic Program Guides
better **precision** [Musto12]

Social data from Facebook and Twitter
better **serendipity** [Narducci13]
i.e. more unexpected and interesting recommendations

Multimedia recommendation::TED lectures
better results in a **cold-start** setting [Pappas14]

[Musto12] C. Musto, F. Narducci, P. Lops, G. Semeraro, M. de Gemmis, M. Barbieri, J. H. M. Korst, V. Pronk, and R. Clout. Enhanced semantic tv-show representation for personalized electronic program guides. UMAP 2012, pp. 188–199. Springer, 2012

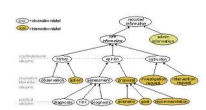
[Narducci13] F. Narducci, C. Musto, G. Semeraro, P. Lops, and M. de Gemmis. Leveraging Encyclopedic Knowledge for Transparent and Serendipitous User Profiles. UMAP 2013, pp. 350-352.

[Pappas14] N. Pappas and A. Popescu-Belis. Combining content with user preferences for non-fiction multimedia recommendation: A study on ted lectures. Multimedia Tools and Applications, 2014.

Top-down approaches

use of

ontological knowledge from simple **linguistic ontologies** to more complex domain-specific ones

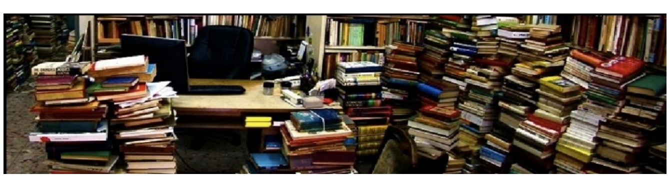


unstructured encyclopedic knowledge sources, such as **Wikipedia**

Linked Open Data cloud

See:

Linked Open Data-enabled Strategies for Top-N Recommendations
C. Musto, P. Basile, P. Lops, M. de Gemmis, G. Semeraro



alternative representation based on
distributional models

insight

the analysis of **large corpora of textual data** allows to infer information about the usage (**meaning**) of the terms

Bottom-up approaches

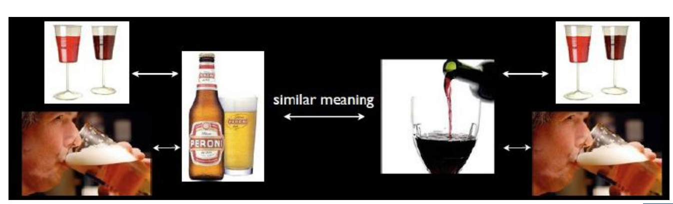
Bottom-up approaches

Distributional models

Distributional Hypothesis

"Meaning is its use"

words that share similar contexts (usages) share similar meaning



Distributional models

term-context matrix

		c1	c2	c3	c4	c5	c6	c7	c8	c9
t1		✔		✔	✔					✔
t2		✔		✔			✔			✔
t3		✔			✔					✔
t4			✔				✔	✔	✔	

different context granularities
document, paragraph, sentence, words

Distributional models

similarity between terms

		c1	c2	c3	c4	c5	c6	c7	c8	c9
t1		✔		✔	✔					✔
t2		✔		✔			✔			✔
t3		✔			✔					✔
t4			✔				✔	✔	✔	

beer vs. glass **good overlap**

Distributional models

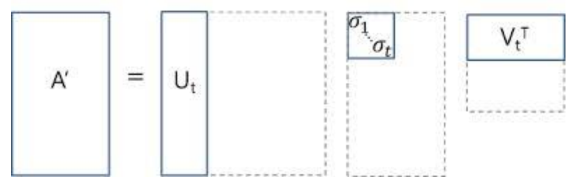
similarity between terms

		c1	c2	c3	c4	c5	c6	c7	c8	c9
t1		✔		✔	✔					✔
t2		✔		✔			✔			✔
t3		✔			✔					✔
t4			✔				✔	✔	✔	

beer vs. spoon **no overlap**

Dimensionality reduction

Truncated Singular Value Decomposition



Induces **higher-order (paradigmatic)** relations through the truncated SVD

Dimensionality reduction

Singular Value Decomposition

PROBLEM

the huge co-occurrence matrix

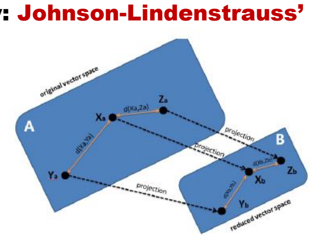
SOLUTION

don't build the huge co-occurrence matrix!

Dimensionality reduction

Random Indexing

theory: Johnson-Lindenstrauss' lemma



$B^{m,k} \approx A^{m,n} R^{n,k} \quad k << n$

distances between the points in the reduced space **approximately preserved**

M. Sahlgren. The Word-Space Model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations between Words in High-dimensional Vector Spaces. PhD thesis, Stockholm University, 2006.

Random Indexing

Algorithm

Step 1 - definition of the context granularity:

Document? Paragraph? Sentence? Word?

Step 2 – building the random matrix R

each **'context'** (e.g. sentence) is assigned a **context vector**

- ✓ dimension = **k**
- ✓ allowed values = **{-1, 0, +1}**
- ✓ small # of non-zero elements, i.e. **spase vectors**
- ✓ values distributed in a **random way**

Random Indexing

Context vectors of dimension k = 8

r_1	0,	0,	-1,	1,	0,	0,	0,	0
r_2	1,	0,	0,	0,	0,	0,	0,	-1
r_3	0,	0,	0,	0,	0,	-1,	1,	0
r_4	-1,	1,	0,	0,	0,	0,	0,	0
r_5	1,	0,	0,	-1,	1,	0,	0,	0
r_n	...							

Random Indexing

Algorithm

Step 3 – building the reduced space B

the **vector space representation** of a **term t** obtained by **combining** the **random vectors** of the **context in which it occurs in**

r_1	0,	0,	-1,	1,	0,	0,	0,	0
r_2	1,	0,	0,	0,	0,	0,	0,	-1
r_3	0,	0,	0,	0,	0,	-1,	1,	0
r_4	-1,	1,	0,	0,	0,	0,	0,	0
r_5	1,	0,	0,	-1,	1,	0,	0,	0
...								
r_n	...							

t1 ∈ {c1, c2, c5}

Random Indexing

Algorithm

Step 3 – building the reduced space B

r_1	0,	0,	-1,	1,	0,	0,	0,	0
r_2	1,	0,	0,	0,	0,	0,	0,	-1
r_3	0,	0,	0,	0,	0,	-1,	1,	0
r_4	-1,	1,	0,	0,	0,	0,	0,	0
r_5	1,	0,	0,	-1,	1,	0,	0,	0
...								
r_n	...							

t1 ∈ {c1, c2, c5}

r_1	0,	0,	-1,	1,	0,	0,	0,	0
r_2	1,	0,	0,	0,	0,	0,	0,	-1
r_5	1,	0,	0,	-1,	1,	0,	0,	0
t1	2,	0,	-1,	0,	1,	0,	0,	-1

Output: **WordSpace**

Random Indexing

Algorithm

Step 4 – building the document space

the **vector space representation** of a **document d** obtained by **combining** the **vector space representation** of the **terms that occur in the document**

Output: **DocSpace**

WordSpace and DocSpace

WordSpace

	c_1	c_2	c_3	c_4	...	c_k
t_1						
t_2						
t_3						
t_4						
...						
t_m						

DocSpace

	c_1	c_2	c_3	c_4	...	c_k
d_1						
d_2						
d_3						
d_4						
...						
d_n						

k is a simple parameter of the model

Uniform representation

eVSM
novel recommendation framework

distributional models for representing **semantics**

dimensionality reduction with **random indexing**

user profiles as **combination** of **vectors of items** the **user liked**

recommendation as **similarity** between **items** and the **user profile** in the **DocSpace**

Cataldo Musto: Enhanced vector space models for content-based recommender systems. RecSys 2010: 361-364

eVSM: some results

Movie domain

MovieLens::content crawled from Wikipedia
better average precision
than classical content-based algorithms [Musto11]

IMDB::contextual eVSM
better **precision** and **recall**
than state-of-the-art approaches based on CF [Musto13]

IMDB::contextual eVSM + entity linking
better **precision** and **recall**
than state-of-the-art approaches based on CF [Musto14]

[Musto11] C. Musto, G. Semeraro, P. Lops, and M. de Gemmis. Random indexing and negative user preferences for enhancing content-based recommender systems. In EC-Web 2011, vol. 85 of LNBIP, pp. 270–281.

[Musto13] C. Musto, G. Semeraro, P. Lops, and M. de Gemmis. Contextual evsm: A content-based context-aware recommendation framework based on distributional semantics. In EC-Web 2013, vol. 152 of LNBIP, pp. 125–136.

[Musto14] C. Musto, G. Semeraro, P. Lops, and M. de Gemmis. Combining distributional semantics and entity linking for context-aware content-based recommendation. In UMAP 2014, vol. 8538 of LNCS, pp. 381–392.

Top-down vs. Bottom-up

Approach		Transp.	Coverage of topics	NLP effort
Top-down	Ontologies		LIMITED	LOW
	Encycl. Knowledge		WIDE	HIGH
Bottom-up	Random Indexing		--	LOW

NLP methods to improve CBRSs

semantics to capture the meaning of content/user needs

overcoming **limited content analysis**

overcoming **overspecialization**

Conclusions