Università degli Studi di Modena e Reggio Emilia
Facoltà di Ingegneria – Sede di Modena
Corso di Laurea Specialistica in Ingegneria Informatica

Development and Application of Semantic Web Technologies in the Area of Personalized Content Distribution

Relatore:
Prof.ssa Sonia Bergamaschi

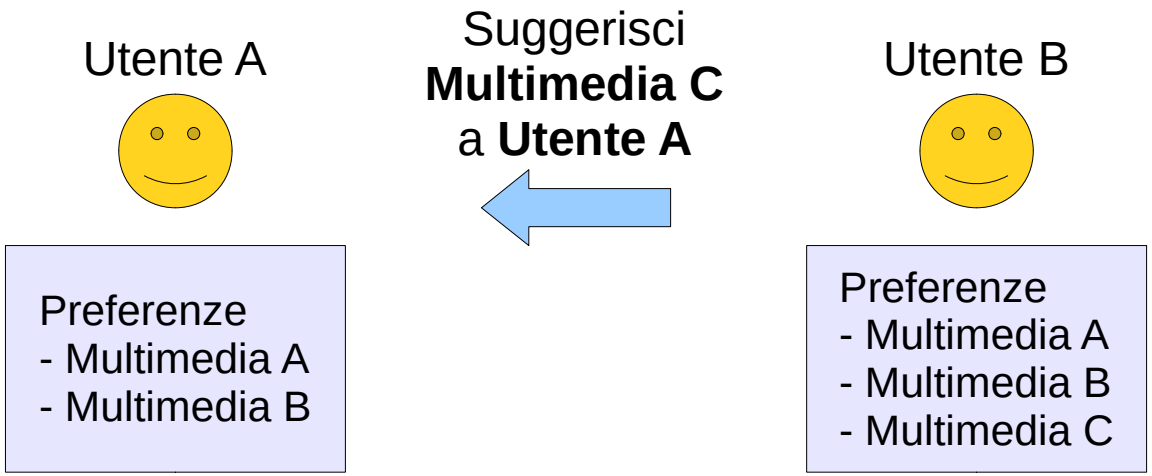
Candidato:
Tania Farinella

Anno accademico 2010 - 2011

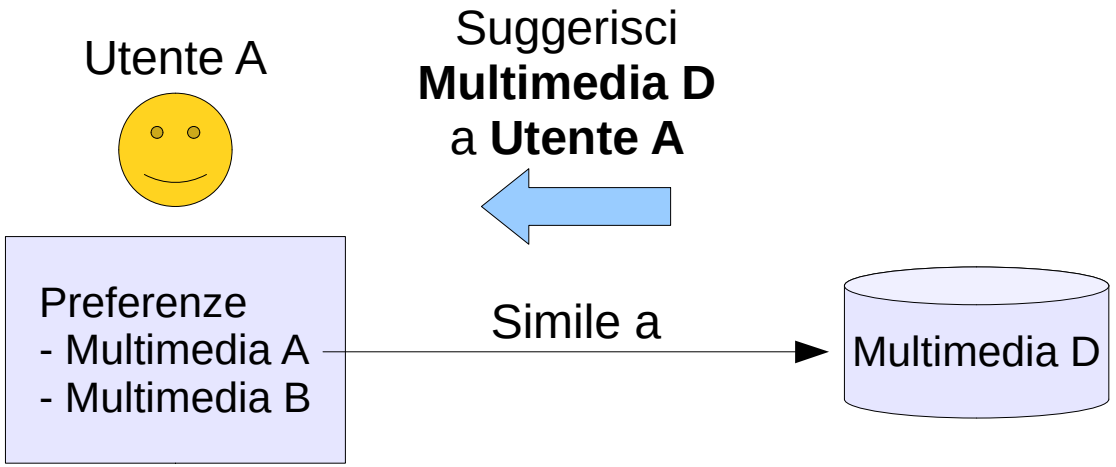
Outline

1. Sistemi di Recommendation
2. Importazione Dati
3. Confronto Trame
4. Confronto Multimedia
5. Conclusioni e Sviluppi Futuri

Collaborative Filtering



Content-based



Risorse Esterne Utilizzate

Internet Movie Database (IMDb)

- Dettagliato e completo
- Licenza non commerciale

DBpedia

- Enciclopedico
- RDF
- Ridistribuzione libera

DBpedia Deutschland

- Tedesco

The Open Movie Database (TMDb)

- API
- Stesse trame di DBpedia

NoSQL

Schema libero

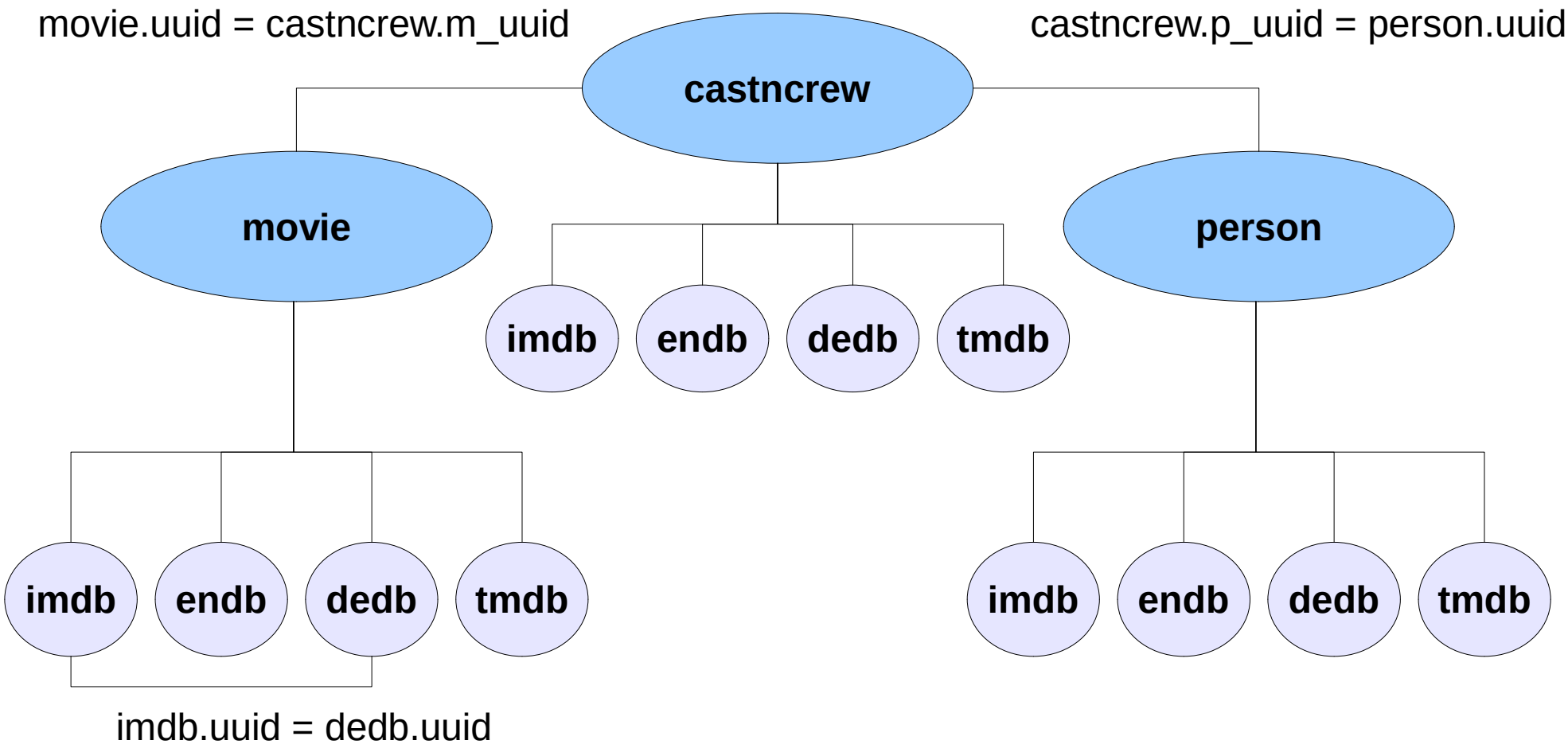
- Collection (tabelle)
- Document (tuple)

Document Oriented

Query language

- No Join ⟹ Attributi Condivisi

Database Locale



Tecniche testate

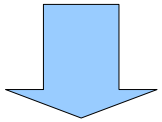
- Term Frequency-Inverse Document Frequency
- Log Entropy
- Latent Semantic Analysis

Cosine Similarity

$$coseno\left(v_1, v_2\right) = \frac{\sum_k \left(v_1[k] \cdot v_2[k]\right)}{\sqrt{\sum_k v_1[k]^2} \cdot \sqrt{\sum_k v_2[k]^2}}$$

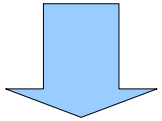
Modello vettoriale

Trama



Keyword 1	**Keyword 2**	**Keyword 3**	**...**	**Keyword 1.500.000**
Peso 1	Peso 2	Peso 3	...	Peso 1.500.000

Training set



	Keyword 1	**Keyword 2**	**...**	**Keyword 1.500.000**
Trama 1	Peso 1,1	Peso 1,2	...	Peso 1,1.500.000
Trama 2	Peso 2,1	Peso 2,2	...	Peso 2,1.500.00
...	...	...	...	...
Trama 200.000	Peso 200.000,1	Peso 200.000,2	...	Peso 200.000,1.500.000

Term Frequency – Inverse Document Frequency (TF-IDF)

$$peso_{tf-idf} = \frac{tf \cdot idf}{norm(v)}$$

Log Entropy (LOG)

$$peso_{\log} = \log(tf + 1) \cdot \left(1 + \frac{\sum_{l=1}^{N} P(v_l, k) \cdot \log\left(P(v_l, k)\right)}{\log(N)} \right)$$

	Francia	Germania	paese	stato	UE	...
La Francia è il più grande paese dell'UE.	0.5	0.0	0.3	0.0	0.7	...
L'UE è costituita da 27 stati.	0.0	0.0	0.0	0.3	0.6	...
La Germania è uno stato federale.	0.0	0.5	0.0	0.4	0.0	...

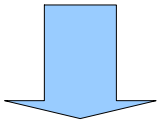
Latent Semantic Analysis (LSA)

- Sinonimia
- Polisemia

	Francia	Germania	paese	stato	UE	...
La Francia è il più grande paese dell'UE.	0.5	0.1	0.3	0.1	0.7	...
L'UE è costituita da 27 stati.	0.1	0.1	0.1	0.3	0.6	...
La Germania è uno stato federale.	0.1	0.5	0.1	0.4	0.1	...

LSA vs TF-IDF

Inception



TF - IDF	LSA
1. Cobb	1. The Dream Team with Annabelle and Michael
2. Somewhere in Georgia	2. Rainbow's Children
3. Firecreek	3. In Pursuit of a Dream
4. House IV	4. Persistence of her Memory
5. Whiplash	5. Twenty-Seven Stories

Efficienza

Problematiche

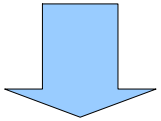
- LSA (decomposizione ai valori singolari)
- Confronto trama – database

Soluzioni

- Gensim
- Riduzione dimensioni matrice

Titolo, Genere, Attori, Produzione, Anno di distribuzione

The Matrix



Solo Trama	Più Attributi
1. Plug & Pray	1. The Matrix Reloaded
2. Die Millennium-Katastrophe – Computer-Crash 2000	2. The Matrix Revolutions
3. Computer Warriors	3. Nezi: The Night of the Crazy Screws
4. Colossus: The Forbin Project	4. In the Realm of the Hackers
5. The KGB, the Computer and Me	5. Plug & Pray

Obiettivi raggiunti

- Confronto trama – database
  - Efficace: LSA
  - Efficiente: 200.000 confronti in meno di un minuto
- Recommendation System senza analisi profilo utente

Sviluppi futuri

- Integrazione ulteriori risorse esterne
- Modello utente (rating)
- Profilo utente (età, genere)
- Navigazione delle informazioni