Implicit News Recommendation Based on User Interest Models and Multimodal Content Analysis

Riccardo Di Massa[1], Maurizio Montagnuolo[2], *Alberto Messina[2]*

[1] Politecnico di Torino

[2]RAI – Radiotelevisione Italiana – Centre for Research and Technological Innovation

AIEMPro10: 3rd ACM Workshop on Automated Information Extraction in Media Production
ACM Multimedia 2010 - Florence, 29th October 2010

Agenda

- **Motivations**
- **Related Work**
- **System Architecture**
- **Experimental design and discussion**
- **Illustrative example**
- **Conclusions and future work**

Motivations for News Recommendation

- **Explosive growth of available informative assets**
  - Professional contents, e.g. Newspapers, press agencies, Radio and TV newscasts
  - Amateur contents, e.g. UGC, social networks, personal blogs

- **Heterogeneity of delivery networks, e.g. the Internet, radio, TV, print media, content and production formats**
- **Finding the most interesting information w.r.t. the user's needs may be waste of time**

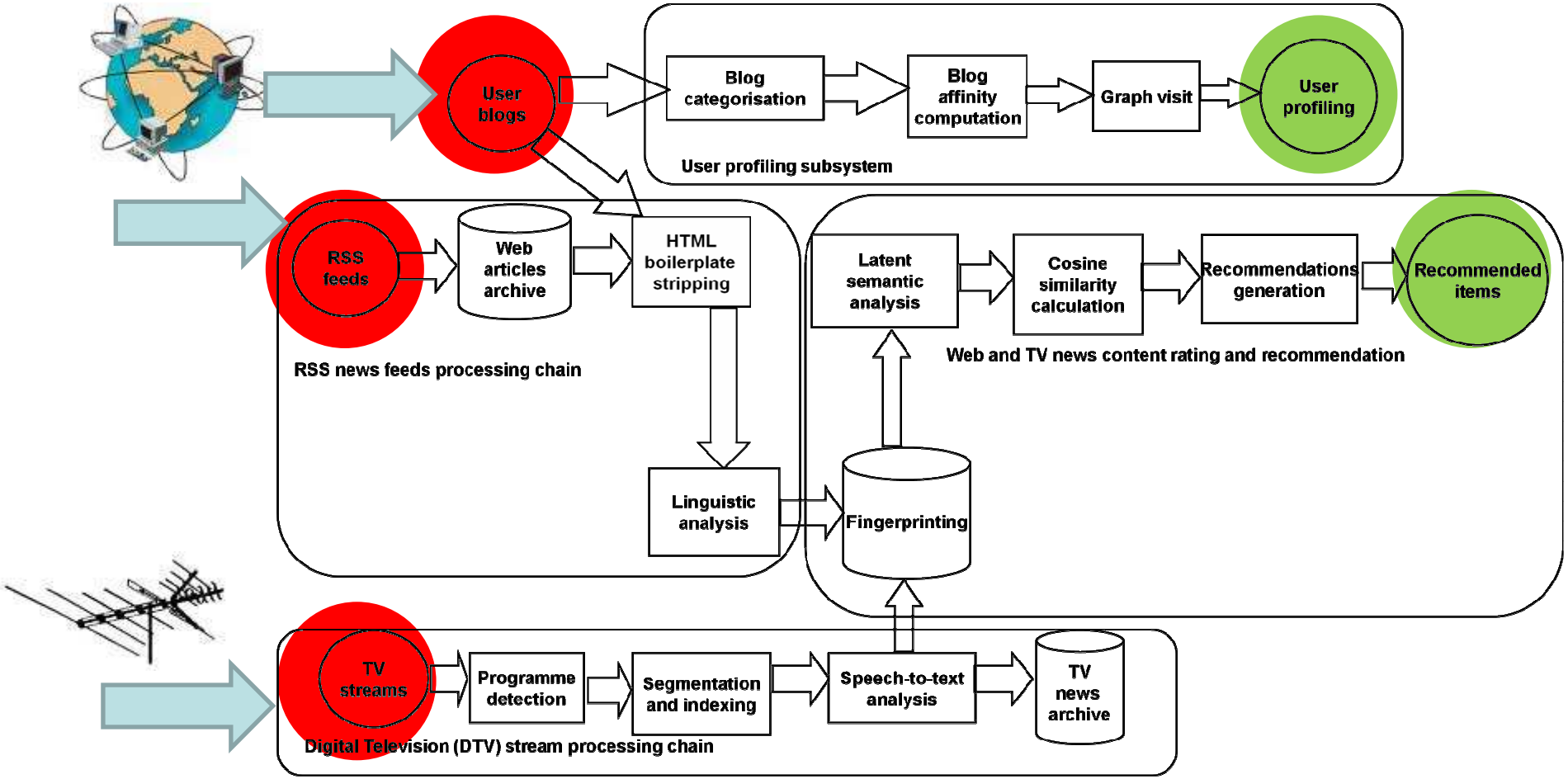
Related Work

- **Item recommendation is not new and mass-usage systems already exist**
  - Collaborative filtering: similar users share similar attitudes
  - Content-based filtering: similar contents are of interest for a user
  - E.g. News recommendation includes Newswine, Newser, Google News
- **Largely dependent on the availability of users' profiles and/or text context descriptors**
- **Mono-modality, i.e. availability of only a single publication domain, e.g. either TV or the Internet**
- **Feeble integration among professional producers and user-generated content**

Goals and Innovations of this Work

- **This work present a complete framework that provides to each user a daily flow of multi-modal news information personalised according to his/her personal interests**
- **Multimodality=multi-media + multi-source**
- **The user's interests are derived by analysing his/her personal blogs, assuming that what he/she wrote can be interpreted as a good approximation of his/her newsworthiness**
- **Some users may simply decide to select feeds of his/her interest instead of the personal blog**
- **The recommendation process is transparent to the user, who is not requested to provide any personal information, neither to adopt any particular formal language in his/her writings**

System Architecture



Digital Television Stream Processing

Performed by ANTS (Automatic Newscast Transcription System), which provides a set of tools for automated news segmentation, classification, indexing and retrieval

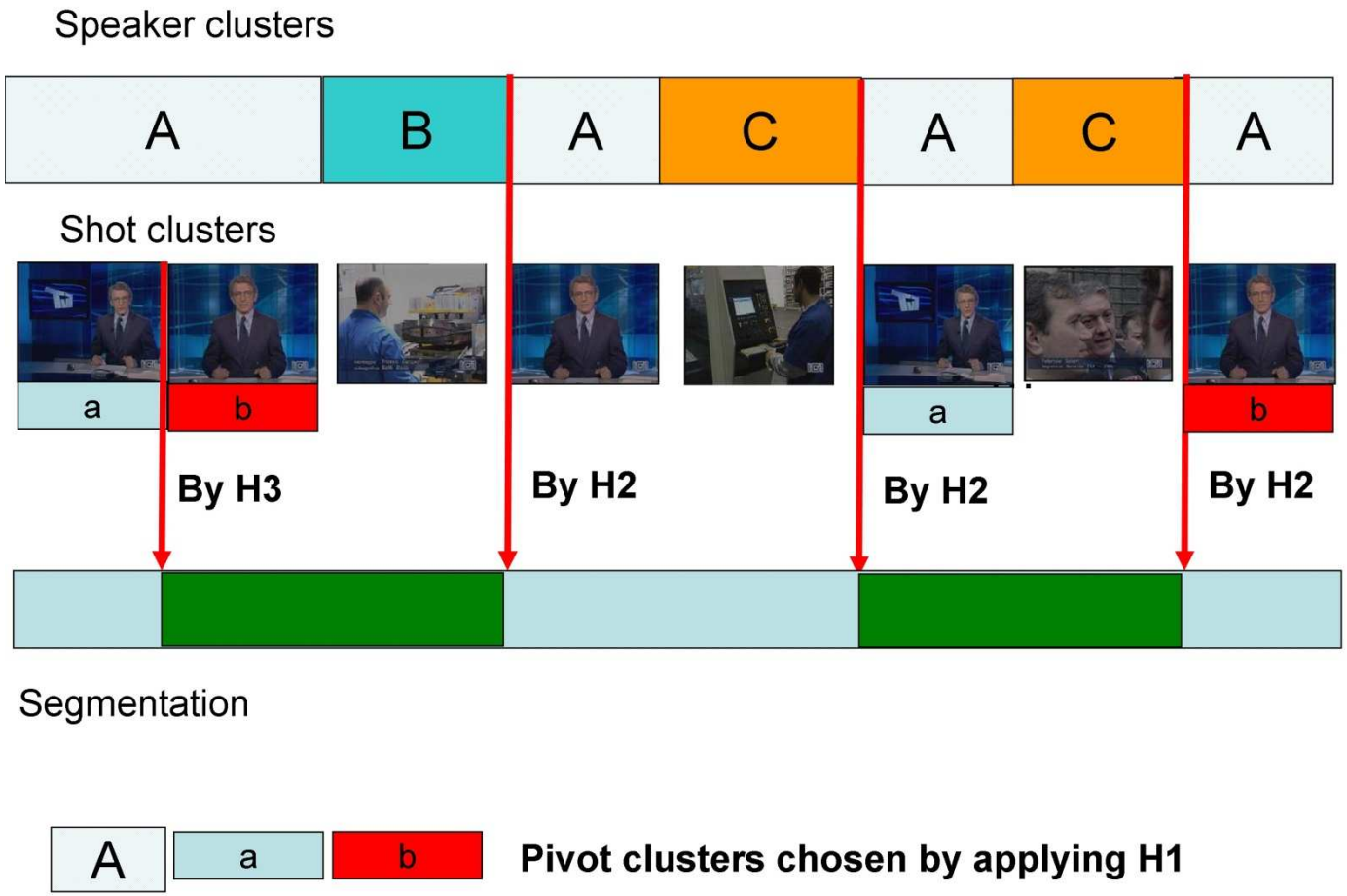
Composed of three main modules

- Programme detection detects the start/end positions of newscasts from the acquired DTV streams
- News story segmentation performs segmentation of the acquired programmes into elementary news stories
- Speech to text analysis extracts text and semantics (i.e., categories, named entities) from the speech content of each story

Programme detection and segmentation

- **Visual pattern matching algorithm for programme detection**
  - Training on detection starting and ending jingles
  - Shot/feature selection based on ranking of Bhattacharyya distances of shots' feature distributions w.r.t. reference population
- **Three-layered heuristics framework for segmentation**
  - H1: Most frequent speaker==anchorman
  - H2: Story boundaries are at anchorman appearances
  - H3: Shot changes during anchorman's speech are story boundaries

Programme detection and segmentation



Segmentation, classification, indexing

- **Once detected, television news items are transcribed, classified, and indexed**
- **Transcription of spoken text is done using an engine optimised for Italian newscasts**
- **Classification is done towards a set of standard categories (RAI) using a Bayesian classifier**
- **Indexing is done by Lucene**

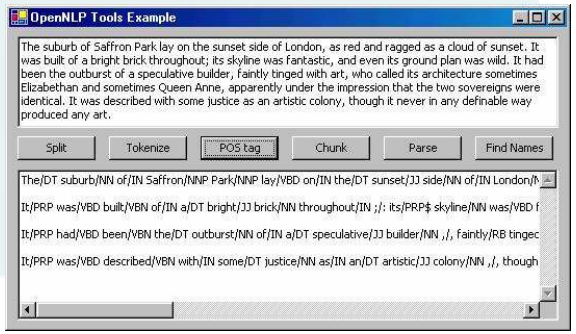
RSS News Stream Processing

Boilerplate stripping extracts the relevant text from the processed Web pages

- E.g. HTML tags, visual formatting, tables, navigation links
- Based on PotaModule, a Perl script available at the SSLMIT (University of Bologna)

Linguistic analysis selects the linguistically interesting parts of text within the extracted texts

- Based on OpenNLP, a set of APIs for Natural Language Processing
- Performed in four steps
  - Sentence Boundary Detection
  - Sentence Tokenisation
  - Word lemmatisation
  - Part of Speech (POS) Tagging



Web and TV News Content Rating and Recommendation

- **Fingerprinting: the fingerprint is a feature vector of dimension N that identifies the content of the processed resources**
  - Computed for each input news resource, i.e. user's blogs, RSS items and TV news stories
  - Made of a vector of words with associated weights corresponding to the occurrence of the word in the input text
  - Example: "*Rumania snow. The southern half of Rumania has been affected by heavy snowfall and blizzards…".*
    The fingerprint is: *d(I)* = (rumania, snow, half, snowfall, blizzard).
- **Latent semantic analysis maps the fingerprints to a vector space of dimension k << N**
- **The similarity between couple of reduced fingerprints is evaluated by the Cosine similarity**
  - Web and TV news fingerprints with high similarity to the user blog fingerprint are recommended to the user

System Evaluation settings

- **330 blog posts from 9 users' blogs**
  1. *"Reality Blog"* - focuses on reality TV programmes
  2. *"Giornalismo d'altri"* - speaks about journalism and new media
  3. *"San Pietro e dintorni"* - Vatican, Catholicism and religion
  4. *"Mediablog"* - technology, computer science, and the Internet
  5. *"La Deriva"* - politics and social welfare
  6. *"Le Vie dell'Asia"* - Asian events
  7. *"Diritto di Cronaca"* - current events
  8. *"Oltretevere"* - Rome and the Vatican City
  9. *"In Diretta da Bruxelles"* – EU affairs and political and administrative institutions
- **8,000 news stories collected from 550 TV newscats**
  - 7 national broadcast channels
- **10,000 articles from 57 RSS feeds**
  - 12 newspapers and press agencies Web sites

Performance Evaluators

Precision at alpha

$$P@\alpha = \frac{CorrectRecs.}{TotalRecs.}$$

Pseudo Recall

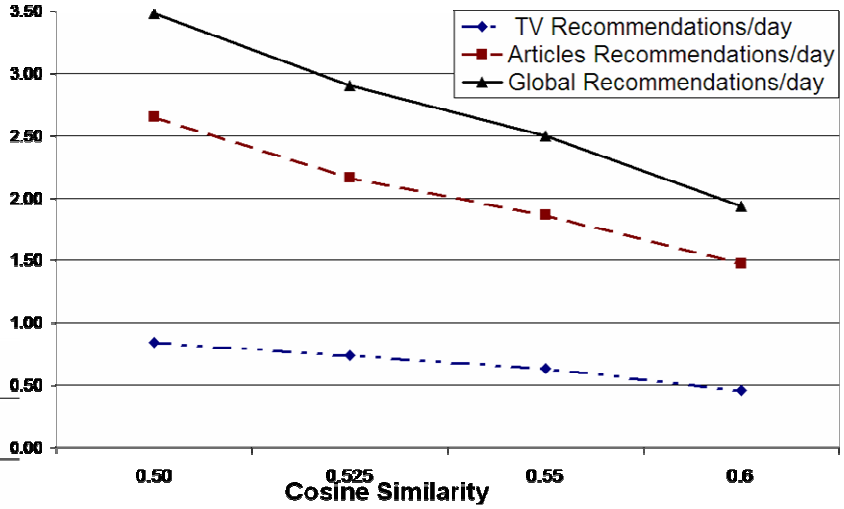
$$PR(\alpha) = 1 - \frac{MissedCorrectRecs.}{CorrectRecs. + MissedCorrectRecs.}$$

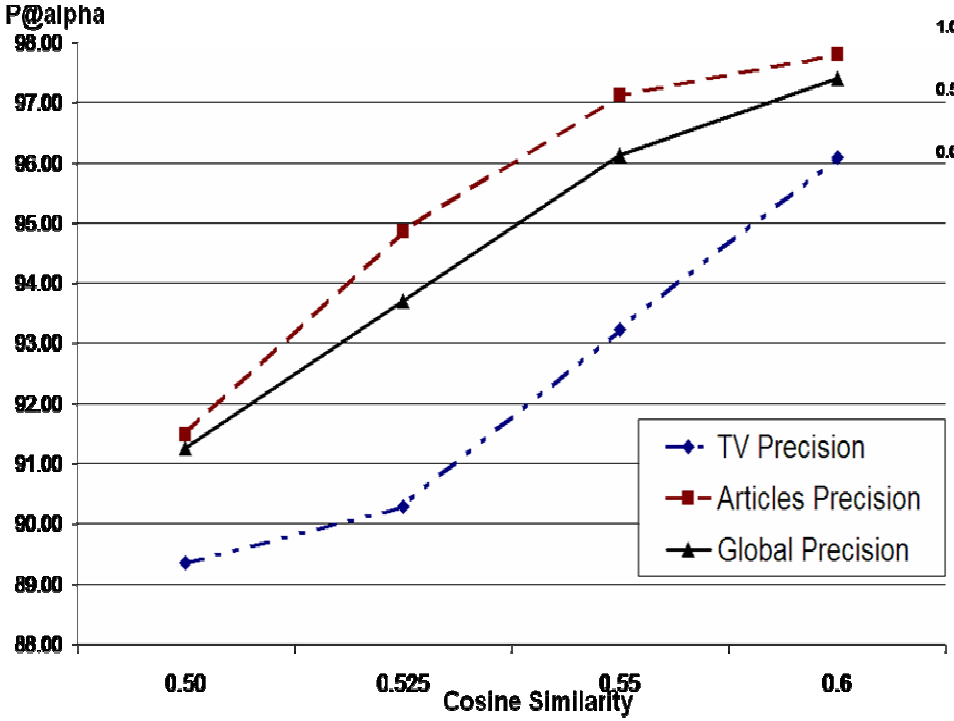
Mean Average Precision

$$MAP = \frac{1}{9}\sum_{b=1}^{9} AP_b = \frac{1}{9}\sum_{b=1}^{9} \frac{1}{R_b}\sum_{i=1}^{R_b} g_b(i)p_b(i)$$

Performance Summary

Threshold (α)	0.5	0.525	0.55	0.6
$P@\alpha$	0.913	0.937	0.963	0.974
$PR(\alpha)$	-	0.856	0.756	0.592
Average daily recommend. per user	3.48	2.90	2.50	1.93





Blog	P@0.5	Average Precision
Reality Blog	0.67	0.59
Giornalismo d'altri	0.63	0.57
San Pietro e dintorni	1.0	1.0
Mediablog	0.78	0.68
La Deriva	0.88	0.82
Le Vie dell'Asia	0.73	0.63
Diritto di Cronaca	0.85	0.80
Oltretevere	1.0	1.0
In Diretta da Bruxelles	0.93	0.89
Mean Average Precision		0.78

User Profiling Subsystem

Collaborative user profiling is performed based on the concept of *semantic affinity*, which is a broad generalisation of the concept of similarity

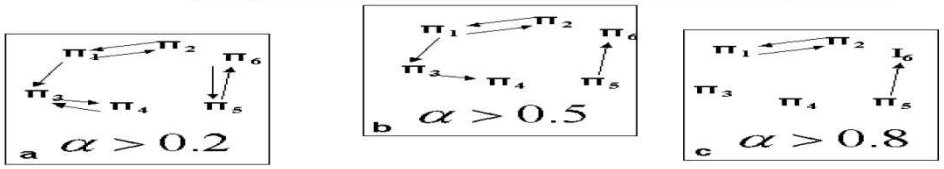
- Partly inspired by human relationships
  - brothers are "similar", (or even equal!)
  - grandpa and his nephews are "affine"
- As such, affinity is also **asymmetrical**

By introducing this generalisation our aim is to define hierarchical relations between users in terms of their affinity in blog posts.

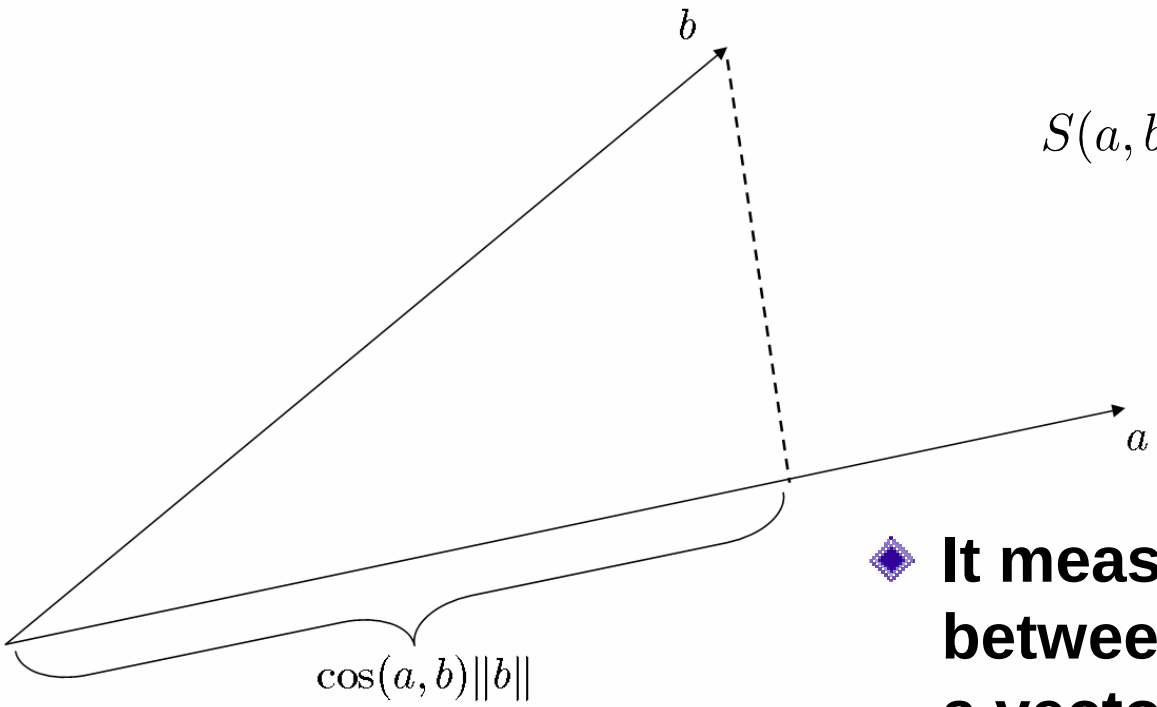
$$S(B_a, B_b) = \frac{\langle \mathbf{c}^{(B_a)}, \mathbf{c}^{(B_b)} \rangle}{\|\mathbf{c}^{(B_a)}\|^2} \quad \mathrm{S}(\mathrm{B}_b, B_a) = \frac{\langle \mathbf{c}^{(B_b)}, \mathbf{c}^{(B_a)} \rangle}{\|\mathbf{c}^{(B_b)}\|^2}$$

Affinities among blogs (and thus among users) can be represented by a disconnected, oriented graph.

	π_1	π_2	π_3	π_4	π_5	π_6
π_1	1	0.82	0.53	0	0	0
π_2	0.85	1	0	0	0	0
π_3	0	0	1	0.6	0	0
π_4	0	0	0.32	1	0	0
π_5	0	0	0	0	1	0.9
π_6	0	0	0	0	0.28	1



The Affinity Concept

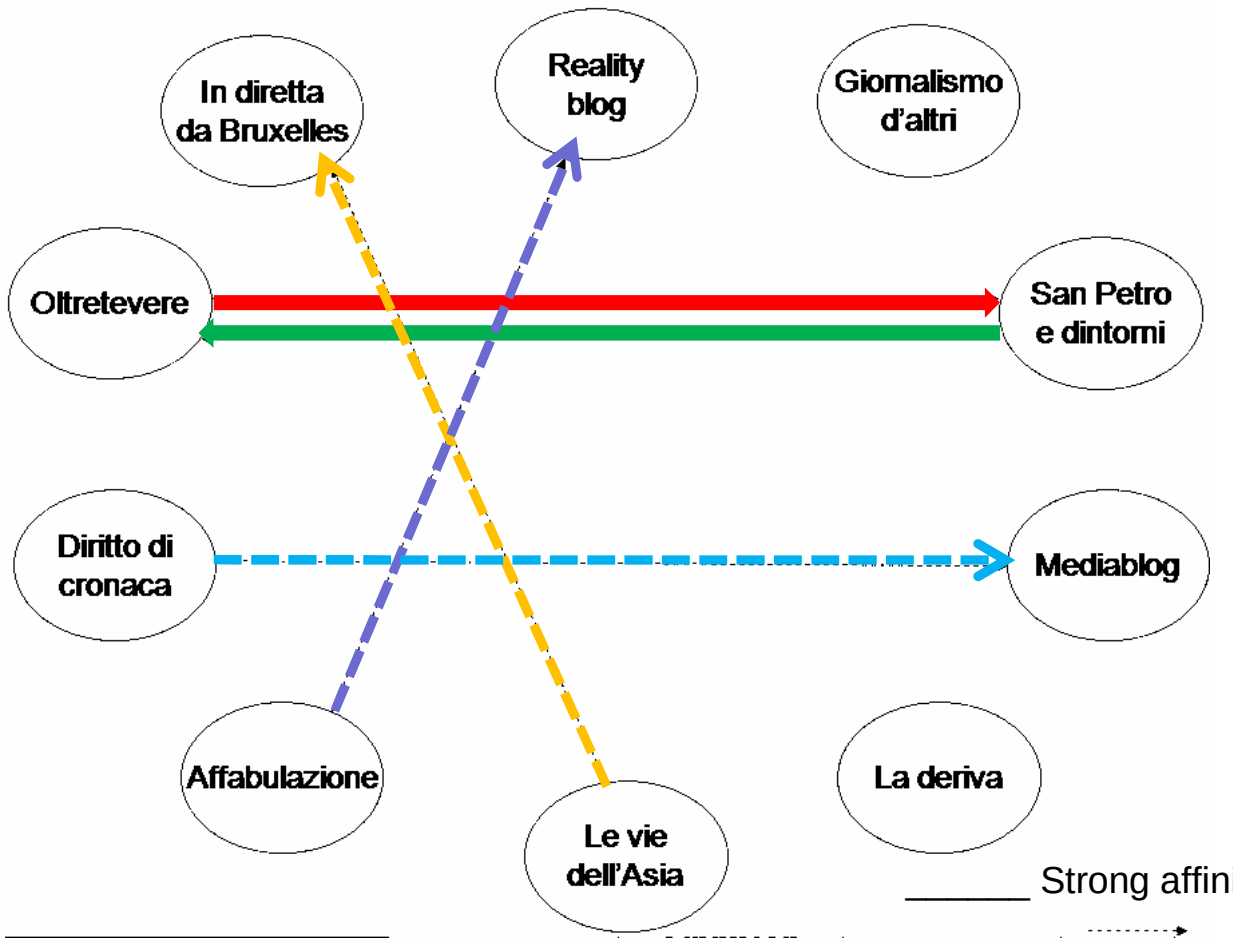


$$S(a,b) = \frac{<a,b>}{\|a\|^2} = \cos(a,b)\frac{\|b\|}{\|a\|}$$

- **It measures the norm-ratio between the projection of a vector a on another vector b, and a**

The defined affinity is a compact index
accounting for angle and size ratio between any two vectors

Empirical Findings on User Profiling



_____ Strong affinity (0.8 < S < 1)

---------- Weak affinity (0.6 < S < 0.8)

Illustrative Example

Profilo :
- Animali
- Cultura Religiosa
- Esteri Paesi Afro-Asiatici
- Esteri Paesi UE
- Scienza
- Sport Calcio
- Tecnologia

User :
- Giuseppe Oddo - Finanza & Potere
- lastampa.it - Diritto di cronaca
- Mario Tedeschini Lalli - Giornalismo d'altri
- rockol.it - Rockol Music News
- wordpress.org - Il Blog di Comuni-Italiani.it

Login

Rockol.it

NewNotizie al concerto di Eugenio Finardi al BlueNote
NewNotizie.it - 2010-10-24 15:18:36 - Score :

Al Blue Note di Milano , tempio milanese del jazz, i concerti generano un'atmosfera unica. Lo spettatore si sente tanto coinvolto da formare, idealmente, una grande orchestra capitanata dall'artista che svetta sul palco. E così è...(continua)

I Pink Floyd tornano in incognito E ci vuole poco per riconoscerli
Il Giornale - 2010-10-24 08:00:00 - Score :

Però si può riassumere con uno slogan: ecco come sarebbe la musica dei Pink Floyd oggi, a quasi quarant’anni da Meddle. Occhio, la musica, e non i testi, perché Metallic Spheres dei The Orb è un...(continua)

Tutti i segreti di Donwood, il pittore «che non c□è»
Il Giornale - 2010-10-24 08:00:00 - Score :

I l 1995 è l□anno in cui in Gran Bretagna è successo davvero di tutto: la consacrazione della young British art nel tempio inglese dell□arte contemporanea, la Royal Academy, con Sensation ; il Turner Prize vinto...(continua)

Mengoni e la musica "Ai fan con amore" il cd 'Re Matto live'
Quotidiano.net - 2010-10-23 15:22:35 - Score :

Milano, 23 ottobre 2010 - Marco coperto , da autunno milanese, al tavolino di un bistrot. Si parla con Mengoni di un cd e dvd live sul “Re Matto Tour”. Semplicemente “Live” (Rca). Registrato all’Arena del...(continua)

I Negramaro con “Sing-hiozzo” già in vetta ai download
NewNotizie.it - 2010-10-23 11:34:35 - Score :

Il brano è on line solo dalla mezzanotte di ieri 22 ottobre ma ha riscosso già un enorme successo: il singolo dei Negramaro e il pre-order del nuovo album "Casa 69" in versione iTunes LP sono...(continua)

TG5 - Servizio Num: 22
Canale5 - 2010-10-24 08:36:19 - Score :

in un anno fa c' e' un modo perfetto non ha ancora trovato il cioccolato perfetta si scopri' extra cacao il nuovo psi fondente di la promozione i giorni della volpe prezzi incredibile anche sull' ... (continua)

TG5 - Servizio Num: 22
Canale5 - 2010-10-22 08:29:51 - Score :

un appello alla pace e' stato punto di partenza per riunire insieme artisti di nazionalita' e di stili diversi l' occasione e' stata fornita dal sinodo dei vescovi per il medio oriente che ha visto esibirsi all' auditorium della ... (continua)

Rai Hyper Media News

Conclusions and Future Work

- **We presented a system for personalised multimodal news recommendation**
- **News items are collected from a variety of sources, including users' blogs, social networks, online newspapers, press sites and TV newscasts, independently from the used writing style**
- **The interests of each user are implicitly deducted by analysing what their posts in public blogs and social network personal pages.**
- **Future work will include**
  - analysing the influence of other performance indicators (such as document length) on the recommendation process
  - investigating efficient visualisation techniques for presenting the recommendations
  - Exploting the inclusion of further blogs, and the extraction of users' opinions.
  - Elaborate more on the collaborative user profiling based on the affinity calculation

For questions and contacts:
Riccardo Di Massa: riccardo.dimassa@gmail.com
Maurizio Montagnuolo: maurizio.montagnuolo@rai.it
Alberto Messina: a.messina@rai.it