

# Studio di modelli di recupero basati su contesti

Dottorato in Informatica ed Applicazioni XXIV Ciclo

## Giuseppe Amodeo

Università de L'Aquila



Fondazione Ugo Bordoni



20-Dicembre-2011

# Introduzione

## Campo di analisi

- Contextual Reasoning
- Information Retrieval
- Text Mining

## Tematiche Affrontate

Integrazione tra

- modelli di recupero
- “contesti” legati a query e utenti

## Casi di studio

In particolare i modelli definiti sono stati applicati a due casi di studio

- Analisi delle opinioni
- Analisi temporali

# Introduzione

## Campo di analisi

- Contextual Reasoning
- Information Retrieval
- Text Mining

## Tematiche Affrontate

Integrazione tra

- modelli di recupero
- “contesti” legati a query e utenti

## Casi di studio

In particolare i modelli definiti sono stati applicati a due casi di studio

- Analisi delle opinioni
- Analisi temporali

# Introduzione

## Campo di analisi

- Contextual Reasoning
- Information Retrieval
- Text Mining

## Tematiche Affrontate

Integrazione tra

- modelli di recupero
- “contesti” legati a query e utenti

## Casi di studio

In particolare i modelli definiti sono stati applicati a due casi di studio

- Analisi delle opinioni
- Analisi temporali

# Formalization of Contextual Reasoning for Information Retrieval

# Contextual Retrieval

L'Information Retrieval storicamente si staglia contro due problematiche

## Uncertainty

*The lack of accuracy in representing the semantics of text and other media*

## Vagueness

*The imprecision of users in expressing their information needs*

## Contextual Retrieval

La ricerca contestuale cerca di superare i limiti dei sistemi di IR tenendo in considerazione le differenze esistenti tra gli utenti così come la pluralità delle informazioni cui un utente può essere interessato.

# Contextual Retrieval

L'Information Retrieval storicamente si staglia contro due problematiche

## Uncertainty

*The lack of accuracy in representing the semantics of text and other media*

## Vagueness

*The imprecision of users in expressing their information needs*

## Contextual Retrieval

La ricerca contestuale cerca di superare i limiti dei sistemi di IR tenendo in considerazione le differenze esistenti tra gli utenti così come la pluralità delle informazioni cui un utente può essere interessato.

# Contextual Retrieval

L'Information Retrieval storicamente si staglia contro due problematiche

## Uncertainty

*The lack of accuracy in representing the semantics of text and other media*

## Vagueness

*The imprecision of users in expressing their information needs*

## Contextual Retrieval

La ricerca contestuale cerca di superare i limiti dei sistemi di IR tenendo in considerazione le differenze esistenti tra gli utenti così come la pluralità delle informazioni cui un utente può essere interessato.

# Operatore Contestuale

Ad ogni documento  $d$  associamo quindi due variabili aleatorie:

$$R = \begin{cases} 1 & \text{se } d \text{ è rilevante} \\ 0 & \text{altrimenti} \end{cases}$$

$$C = \begin{cases} 1 & \text{se } d \text{ rispetta il contesto} \\ 0 & \text{altrimenti} \end{cases}$$

## Probability Ranking Principle

$$O(Y|\vec{d}, \vec{q}) \propto \prod_{i=1}^n \frac{\Pr(x_i|Y=1, \vec{q})}{\Pr(x_i|Y=0, \vec{q})}$$

## Language Model

$$\Pr(d | q, r_c) = \frac{\Pr(q, r_c | d) \cdot \Pr(d)}{\Pr(q, r_c)}$$

### Pubblicazioni

- G. Amodeo. *A Conditional Operator for Contextual Retrieval*, submitted to 3rd Italian Information Retrieval Workshop (IIR), 2012

# Operatore Contestuale

Ad ogni documento  $d$  associamo quindi due variabili aleatorie:

$$R = \begin{cases} 1 & \text{se } d \text{ è rilevante} \\ 0 & \text{altrimenti} \end{cases}$$

$$C = \begin{cases} 1 & \text{se } d \text{ rispetta il contesto} \\ 0 & \text{altrimenti} \end{cases}$$

## Probability Ranking Principle

$$O(Y|\vec{d}, \vec{q}) \propto \prod_{i=1}^n \frac{\Pr(x_i|Y=1, \vec{q})}{\Pr(x_i|Y=0, \vec{q})}$$

## Language Model

$$\Pr(d | q, r_c) = \frac{\Pr(q, r_c | d) \cdot \Pr(d)}{\Pr(q, r_c)}$$

### Pubblicazioni

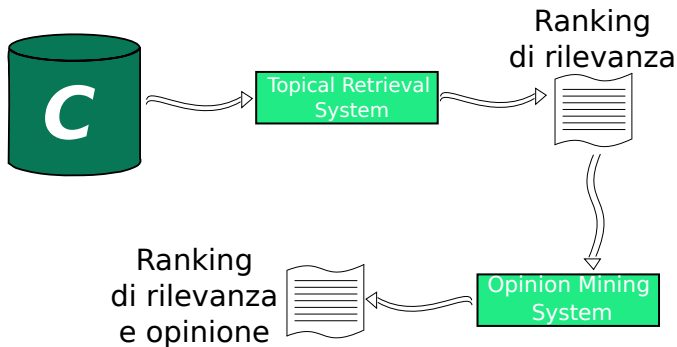
- G. Amodeo. *A Conditional Operator for Contextual Retrieval*, submitted to 3rd Italian Information Retrieval Workshop (IIR), 2012

# OPINION RETRIEVAL

# Introduzione all'Opinion Retrieval

## Opinion Retrieval System

Un sistema di Opinion Retrieval combina un recupero su topic a un'analisi del contenuto sentimentale dei documenti. Scopo di tali sistemi è il recupero e ranking di documenti che siano contemporaneamente rilevanti rispetto ad una query e contenenti opinioni.



## Re-Ranking basato sulle opinioni

Nell'approccio proposto, si vuole stimare la probabilità  $\Pr(q, V \mid d)$  della query  $q$  e del dizionario sentimentale  $V$ :

$$\Pr(q, V \mid d) = \Pr(q \mid d) \cdot \Pr(V \mid d) \propto \frac{\text{Score}_t(d, q)}{r_Y(d)}$$

dove  $\text{Score}_t(d, q)$  è lo score di rilevanza calcolato a partire dalla query e  $r_Y(d)$  è il *rank di opinione* indotto dallo score d'opinione  $\text{Score}_o(d, V)$ .

$k_j$	$\overline{\text{Size}}_j$	$BL_1$		$BL_3$		$BL_4$	
		$\overline{\text{MAP}}_{1,j}$	$\Delta M_{1,j}^{\%}$	$\overline{\text{MAP}}_{3,j}$	$\Delta M_{3,j}^{\%}$	$\overline{\text{MAP}}_{4,j}$	$\Delta M_{4,j}^{\%}$
100	4711.8	0.3024	14.60%	0.3448	7.71%	0.3731	5.31%
1000	1452.8	0.3024	14.60%	<b>0.3449</b>	<b>7.74%</b>	0.3732	5.33%
5000	364.4	<b>0.3027</b>	<b>14.70%</b>	0.3448	7.72%	0.3738	5.52%
10000	147.6	0.3020	14.42%	0.3444	7.59%	<b>0.3743</b>	<b>5.66%</b>
20000	18.4	0.2936	11.26%	0.3402	6.27%	0.3680	3.86%

**Tabella:** Risultato dell'applicazione della tecnica di OR a differenti baseline di rilevanza

## Re-Ranking basato sulle opinioni

Nell'approccio proposto, si vuole stimare la probabilità  $\Pr(q, V \mid d)$  della query  $q$  e del dizionario sentimentale  $V$ :

$$\Pr(q, V \mid d) = \Pr(q \mid d) \cdot \Pr(V \mid d) \propto \frac{\text{Score}_t(d, q)}{r_Y(d)}$$

dove  $\text{Score}_t(d, q)$  è lo score di rilevanza calcolato a partire dalla query e  $r_Y(d)$  è il *rank di opinione* indotto dallo score d'opinione  $\text{Score}_o(d, V)$ .

$k_j$	$\overline{\text{Size}}_j$	$BL_1$		$BL_3$		$BL_4$	
		$\overline{\text{MAP}}_{1,j}$	$\Delta M_{1,j}^{\%}$	$\overline{\text{MAP}}_{3,j}$	$\Delta M_{3,j}^{\%}$	$\overline{\text{MAP}}_{4,j}$	$\Delta M_{4,j}^{\%}$
100	4711.8	0.3024	14.60%	0.3448	7.71%	0.3731	5.31%
1000	1452.8	0.3024	14.60%	<b>0.3449</b>	<b>7.74%</b>	0.3732	5.33%
5000	364.4	<b>0.3027</b>	<b>14.70%</b>	0.3448	7.72%	0.3738	5.52%
10000	147.6	0.3020	14.42%	0.3444	7.59%	<b>0.3743</b>	<b>5.66%</b>
20000	18.4	0.2936	11.26%	0.3402	6.27%	0.3680	3.86%

**Tabella:** Risultato dell'applicazione della tecnica di OR a differenti baseline di rilevanza

# Conclusioni

## Pubblicazioni

- G. Amati, G. Amodeo, M. Bianchi, C. Gaibisso, and G. Gambosi. *FUB, IASI-CNR and University of Tor Vergata at TREC 2008 Blog Track*, in the Proceedings of The Seventeenth Text REtrieval Conference (TREC 2008) Proceedings, Voorhees E. M., Buckland L. P. eds., NIST Special Publication, SP 500-277, 2009
- G. Amati, G. Amodeo, M. Bianchi, C. Gaibisso, and G. Gambosi. *A uniform theoretic approach to opinion and information retrieval*, in *Intelligent Information Access*. G. Armano, M. de Gemmis, G. Semeraro, and E. Vargiu (eds.) Studies in Computational Intelligence. Springer, 2010

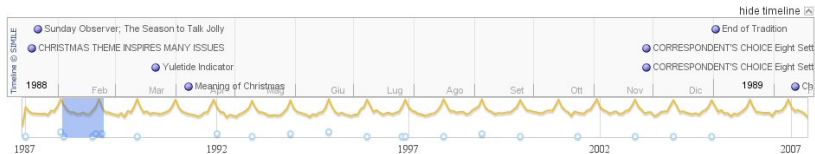
# Analisi della dimensione temporale

## Definizione del Problema

Data una query, ci si chiede come i documenti rilevanti si distribuiscano nel tempo.

### Approccio

Partendo dal query result set di una interrogazione, si costruisce una timeline dei documenti sfruttando il timestamp di pubblicazione dei documenti recuperati.



# Query Expansion Temporale

## Approccio

Definiamo come insieme di documenti pseudo rilevanti quelli che ricadono nel più alto picco di pubblicazioni, considerando gli altri documenti come insieme dei non rilevanti.

$$q_m = \alpha q_0 + \beta \frac{1}{|D_r|} \sum_{d_j \in D_r} d_j - \gamma \frac{1}{|D_{nr}|} \sum_{d_j \in D_{nr}} d_j$$

	Baseline	$\alpha = 1, \beta = 0.9, \gamma = 0.1$	$\alpha = 1, \beta = 1, \gamma = 0$
MAP	0.3067	0.3126	0.3298

# Query Expansion Temporale

## Approccio

Definiamo come insieme di documenti pseudo rilevanti quelli che ricadono nel più alto picco di pubblicazioni, considerando gli altri documenti come insieme dei non rilevanti.

$$q_m = \alpha q_0 + \beta \frac{1}{|D_r|} \sum_{d_j \in D_r} d_j - \gamma \frac{1}{|D_{nr}|} \sum_{d_j \in D_{nr}} d_j$$

	Baseline	$\alpha = 1, \beta = 0.9, \gamma = 0.1$	$\alpha = 1, \beta = 1, \gamma = 0$
MAP	0.3067	0.3126	0.3298

# Re-Ranking su base temporale

## Approccio

In collezioni fortemente legate all'aspetto temporale (**Tweets** o **Blog**), riordinando i documenti recuperati per istante di pubblicazione si può migliorare il ranking. La “freschezza” delle informazioni viene così premiata.

$$p(q|\tau, d) \propto p(\tau|q, d) \cdot p(q|d) = \frac{p(q|d)}{\text{rank}_\tau(d) + B_\tau} \propto \frac{\text{score}_r(d)}{\text{rank}_\tau(d) + B_\tau}$$

	KLIM	KLIM	KLIM30	KLIMRA	KLIMZipf
Ranked by	Relev.	Time	Time	Relev.	Relev.
#tweets Retrieved	1000	1000	30	variable	1000
recip_rank	0.7403	0.7475	0.8219	0.8231	0.7665
$P@30$	0.4395	0.1136	0.4401	0.4476	0.4537

# Re-Ranking su base temporale

## Approccio

In collezioni fortemente legate all'aspetto temporale (**Tweets** o **Blog**), riordinando i documenti recuperati per istante di pubblicazione si può migliorare il ranking. La "freschezza" delle informazioni viene così premiata.

$$p(q|\tau, d) \propto p(\tau|q, d) \cdot p(q|d) = \frac{p(q|d)}{\text{rank}_\tau(d) + B_\tau} \propto \frac{\text{score}_r(d)}{\text{rank}_\tau(d) + B_\tau}$$

	KLIM	KLIM	KLIM30	KLIMRA	KLIMZipf
Ranked by	Relev.	Time	Time	Relev.	Relev.
#tweets Retrieved	1000	1000	30	variable	1000
recip_rank	0.7403	0.7475	0.8219	0.8231	0.7665
P@30	0.4395	0.1136	0.4401	0.4476	0.4537

# Conclusioni

## Pubblicazioni

- G. Amodeo, R. Blanco, U. Brefeld. *Hybrid Models for Future Event Prediction*, in the Proceedings of 20th, ACM International Conference on Information and Knowledge Management, 2011
- G. Amodeo, G. Amati, G. Gambosi. *On relevance, time and query expansion*, in the Proceedings of 20th, ACM International Conference on Information and Knowledge Management, 2011
- G. Amati, G. Amodeo, M. Bianchi, G. Marcone, C. Gaibisso, G. Gambosi, A. Celi, C. De Nicola, M. Flammini. *FUB, IASI-CNR, UNIVAQ at TREC 2011*, to appear in Proceedings of TREC 2011.

# Valutazione dei sistemi di Contextual Retrieval

# Sistema di valutazione

Come valutare i sistemi di CR? Analizziamo il caso dell'Opinion Retrieval

- Vengono utilizzati dei classificatori artificiali utilizzandoli come strumenti "standard" per l'Opinion Mining.

Ranking di rilevanza

Metrica di precisione della Rilevanza

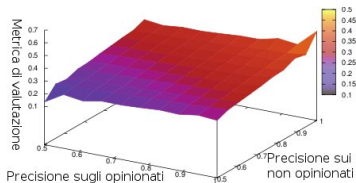
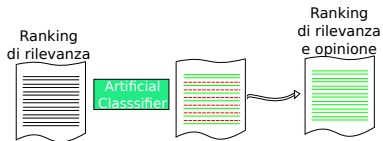
Artificial Classifier

Metrica di precisione di opinione

Metrica di precisione della Rilevanza e di opinione

Ranking di rilevanza e opinione

- si studia come, utilizzando il classificatore artificiale per filtrare i documenti della baseline, vari la qualità finale del sistema al variare della precisione in classificazione



# Conclusioni

## Pubblicazioni

- G. Amati, G. Amodeo, V. Capozio, C. Gaibisso, and G. Gambosi. *Assessing the quality of opinion retrieval systems*, In Proceedings of 1st International Conference on Opinion Mining for Business Intelligence (OMBI), 2010
- G. Amati, G. Amodeo, V. Capozio, C. Gaibisso, and G. Gambosi. *On performance of Topical Opinion Retrieval*, In Proceedings of 33rd ACM International Conference on Special Interest Group on Information Retrieval (SIGIR), 2010
- G. Amati, G. Amodeo, V. Capozio, C. Gaibisso, and G. Gambosi. *A study on the evaluation of opinion retrieval systems*, In Proceedings of 1st Italian Information Retrieval Workshop (IIR), 2010

## Conclusioni

- Definizione formale del Contextual Reasoning per la definizione di modelli di IR
  - Contextual Probability Ranking Principle
  - Contextual Language Model
- Applicazione dei modelli basati su contesti all'Opinion Retrieval
- Applicazione dei modelli basati su contesti al Real-Time Retrieval
- Definizione di un framework di valutazione per i sistemi di Contextual Retrieval

## Contributi

- Intelligent Retrieval in Multimedia Archives (IRMA)
- TV++

Grazie

