Distinguishing the Popularity Between Topics: A System for Up-to-date Opinion Retrieval and Mining in the Web

Nikolaos Pappas,
Georgios Katsimpras,
Efstathios Stamatatos

March 26, 2013

Outline

Motivation

Huge number of user-generated text in the Web

- Many applications based on sentiment analysis
  e.g. brand analysis, marketing effectiveness
- Most approaches focus on fixed collections or certain domains
  e.g. Twitter, Facebook, Blogspot
- Opinion analysis can differ according to the examined web genres
  e.g. articles, blogs, forums

Challenges

→ Collecting domain-independent opinionated texts dynamically from the Web
→ Providing genre-based analysis of opinions
→ Comparing popularity of topics

The proposed system

Synthesis of IR and NLP components:

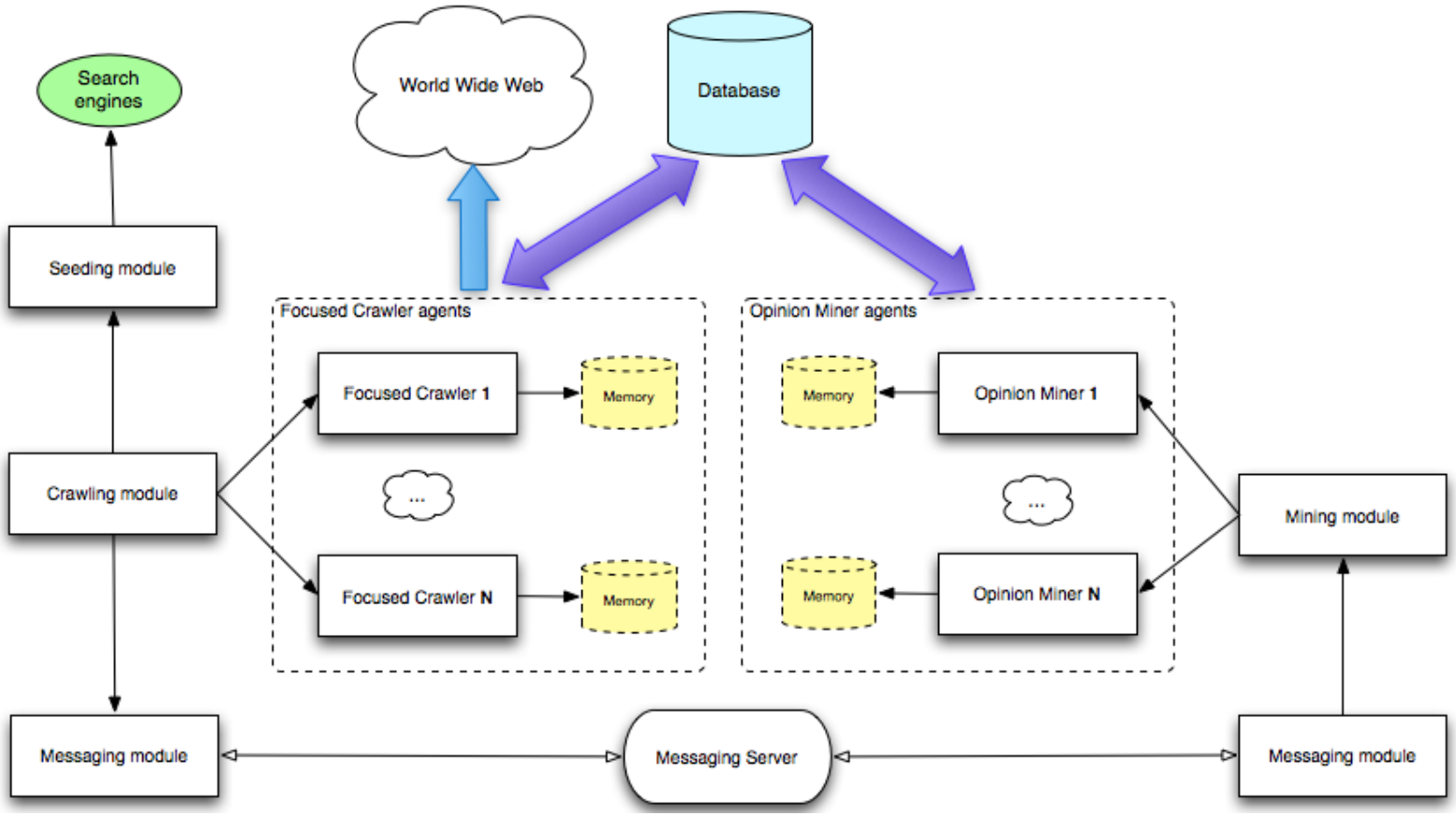
1. Discovery of topic-related documents dynamically from the Web
2. Detection of user-generated content regions inside the related pages
3. Identification of topic-related pages with confidence score
4. Subjectivity and polarity detection on the detected regions

Contributions

→ Up-to-date opinionated text retrieval and mining
→ Genre-based analysis of sentiment
→ Efficient estimation of total sentiment

Evaluation: Qualitative analysis with real-world experiments

Synthesis of IR and NLP components



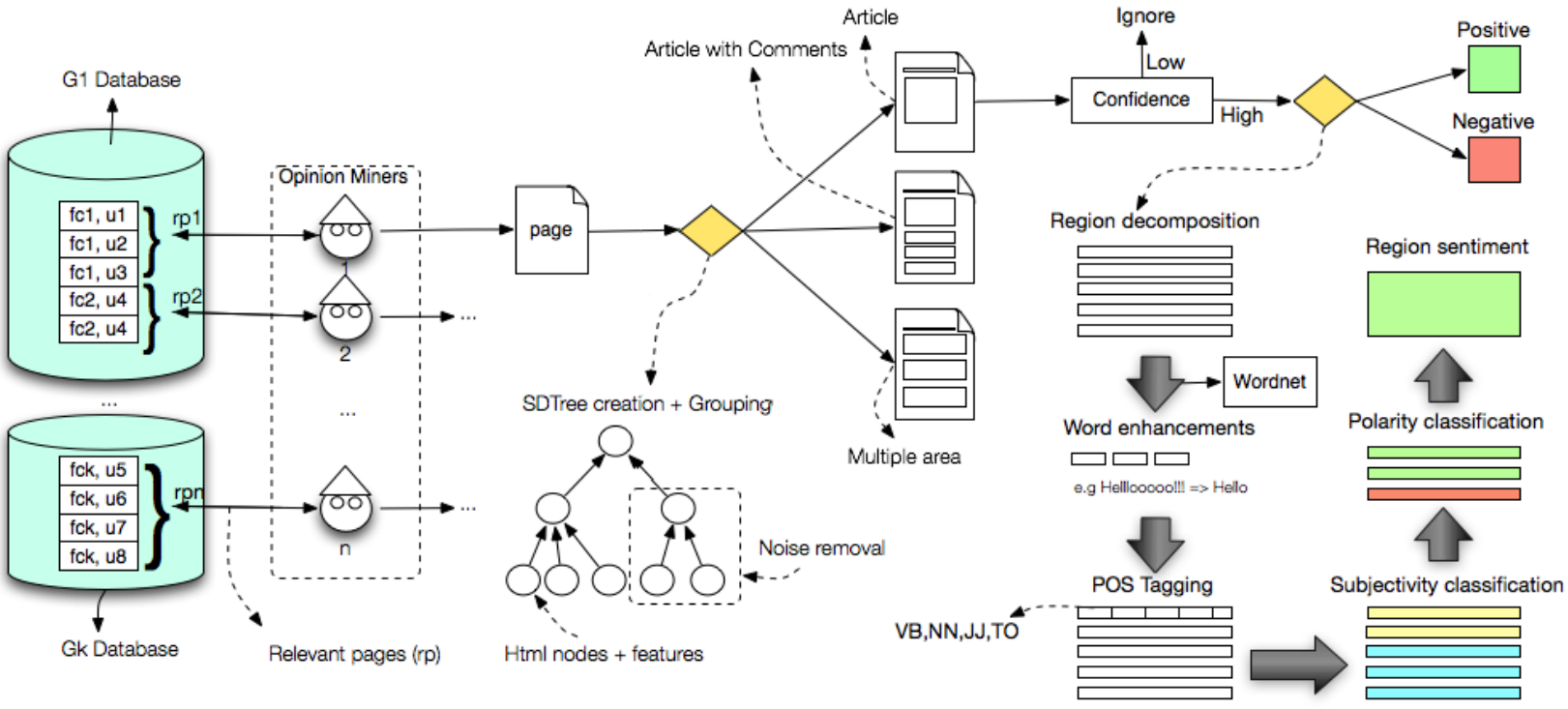
Topic-related document discovery

Given a topic query (keyword form):

- → collection of seed URLs from major search engines
- → topic-(T) and genre-related (G) focused crawling [PKS12a]
- → scoring of unvisited pages using link analysis
  $Linkscore(p) = w_T * Linkscore_T(p) + w_G * Linkscore_G(p)$
- → targeting to web genres (news, blogs, discussions)
  - highly likely to contain opinions

Opinion retrieval and mining

1. Page segmentation and filtering
2. Sentiment analysis



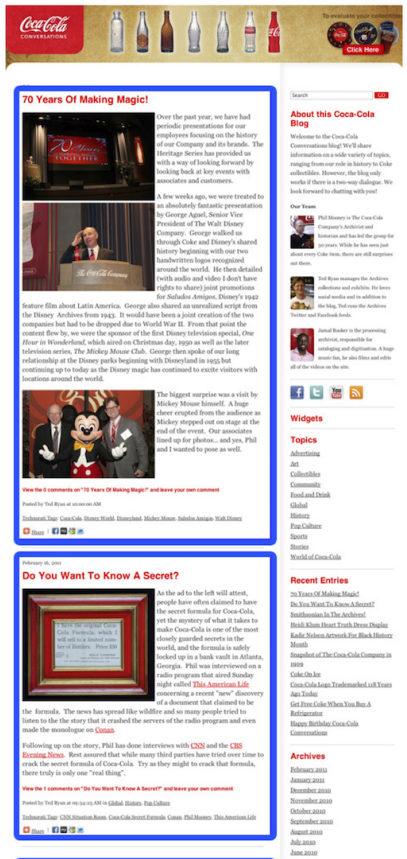
Page segmentation and filtering

Given a web page:

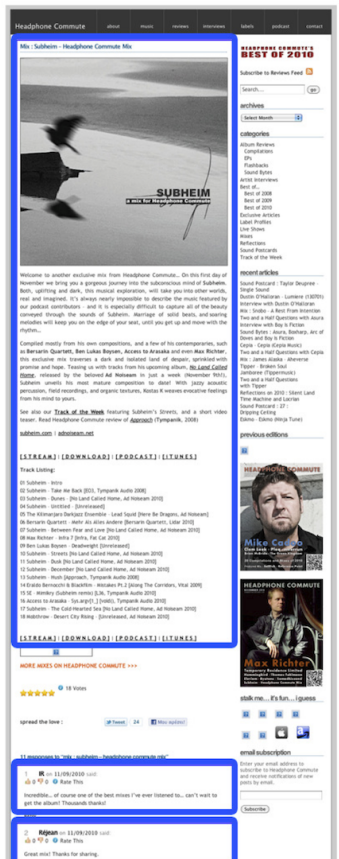
- → segmentation into coherent parts and noise removal (e.g. ads) [PKS12b]
- → rule-based classification and region extraction of three classes
- → confidence of page relevance based on the topic presence (keyword(s)) in the detected regions (weighted linear combination)

(a) Article

(b) Multiple areas



(c) Article with Comments



Sentiment analysis

For each sentence in the detected regions:

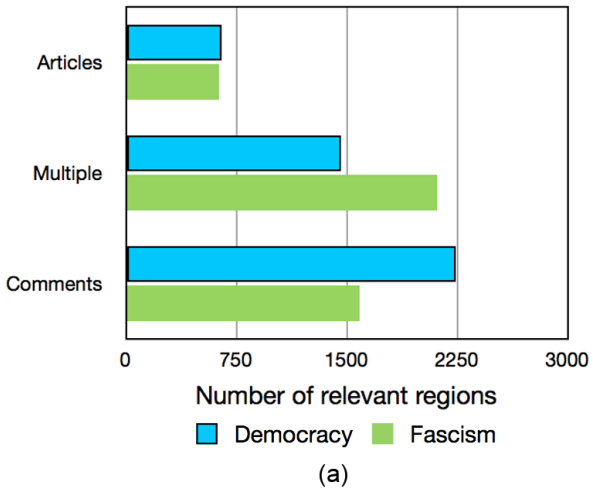
- → subjectivity classification (bootstrap with Pattern-based learner) [RW03]
- → polarity classification (bootstrap with SVM) [WWH05, MW09]
- → total sentiment estimation

$$TotalScore(D) = \sum_{d_j \in D} \left(\sum_{r_{ij} \in d_j} Score(r_{ij}) \right) \in R \quad (1)$$

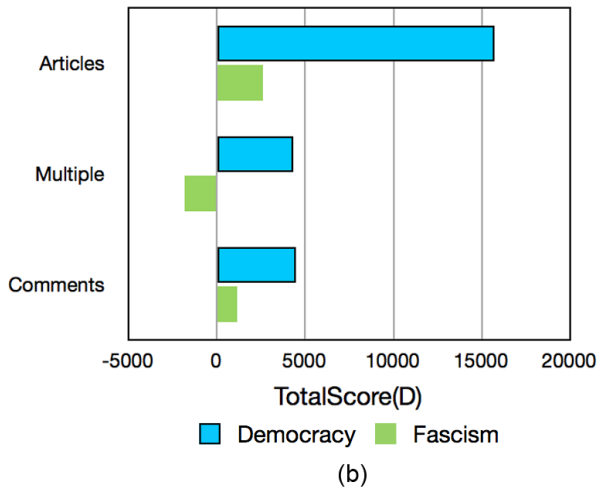
$$NormalizedScore(D) = \sum_{d_j \in D} \left(\sum_{r_{ij} \in d_j} \frac{Score(r_{ij})}{|r_{ij}|} \right) \in R \quad (2)$$

$$SentimentRatio(D) = \frac{|r_{pos}|}{|r_{pos}| + |r_{neg}|} \in [0, 1] \quad (3)$$

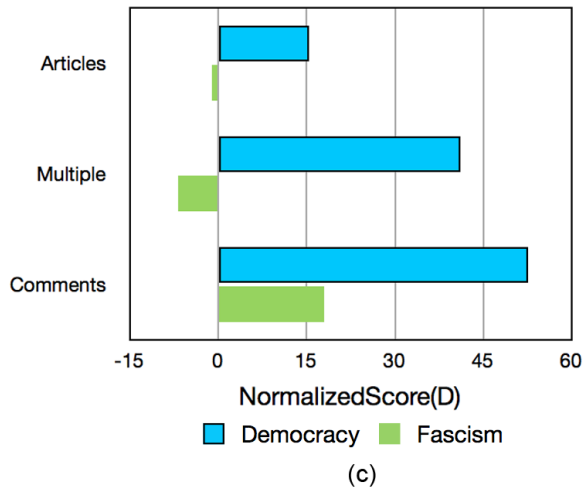
Case study 1: Distinguishing topic popularity



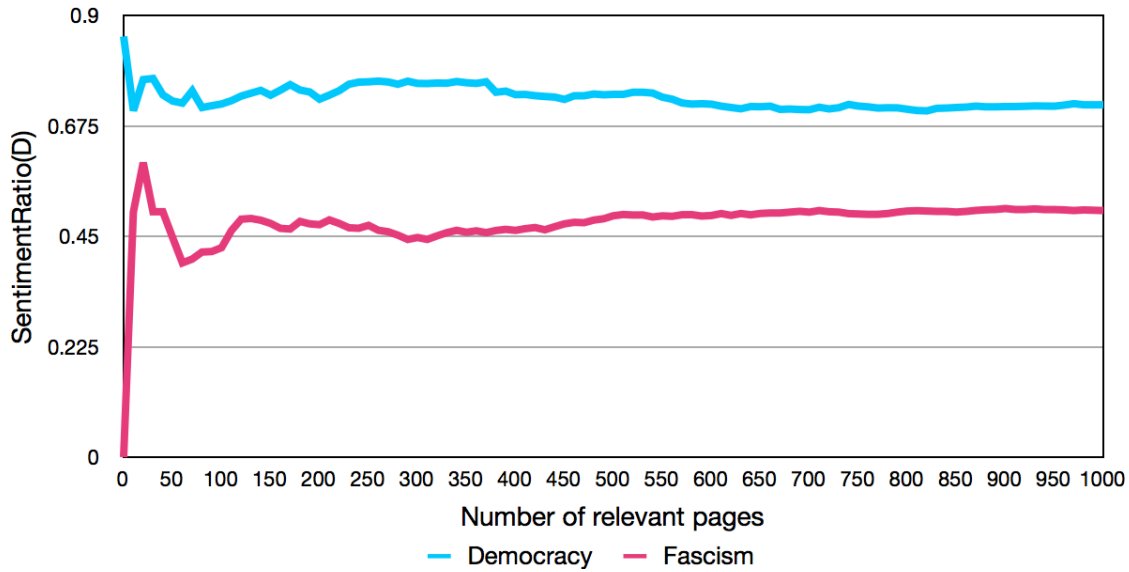
(a)



(b)

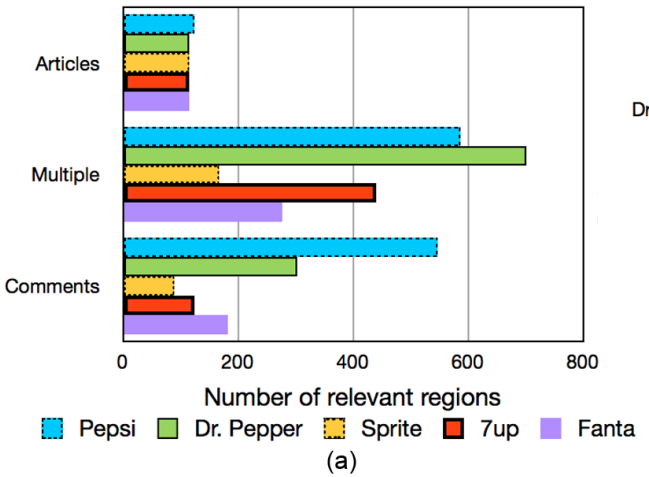


(c)

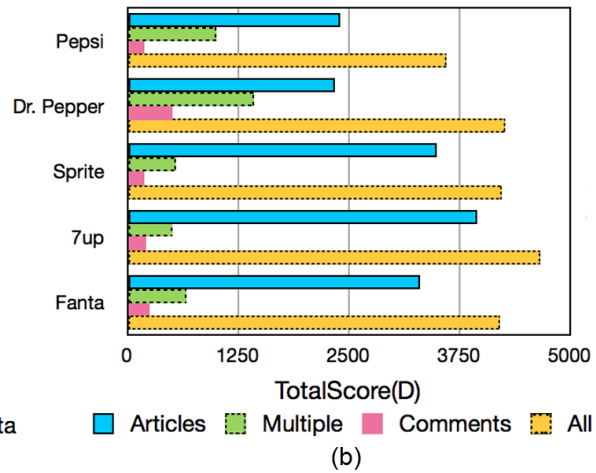


Case study 1: Opinions per detected regions

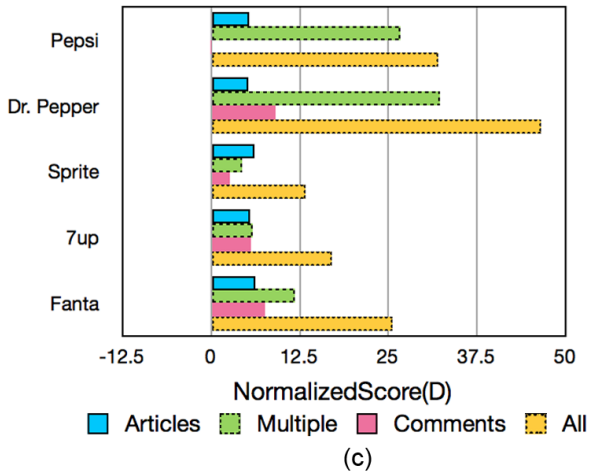
Case study 2: Ranking of topics



(a)



(b)



(c)

Rank	Soft drink	Likes	Talking	Both	TotalScore	NormalizedScore
1_{st}	Dr. Pepper	12,093,912	187,011	12,280,923	7up	Dr. Pepper
2_{nd}	Pepsi	11,835,244	236,105	12,071,349	Dr. Pepper	Pepsi
3_{rd}	Sprite	8,574,563	50,192	8,624,755	Sprite	Fanta
4_{th}	Fanta	2,650,072	84,080	2,734,152	Fanta	7up
5_{th}	7up	785,967	75,996	861,963	Pepsi	Sprite
	IM Client	**Followers**	-	-	**TotalScore**	**NormalizedScore**
1_{st}	Google Talk	405,818	-	-	Google Talk	Google Talk
2_{nd}	Skype	367,385	-	-	Skype	Skype
3_{rd}	MSN	82,896	-	-	MSN	MSN
4_{th}	AOL	14,431	-	-	AOL	ICQ
5_{th}	ICQ	14,138	-	-	ICQ	AOL
				NDCG:	0.841	0.993

Conclusions

- up-to-date discovery of opinionated text for given topics
- genre-aware sentiment analysis of opinions
- real-world studies
  - distinguishing the popularity between topics
  - comparative results for several topics
  - efficient popularity estimation with few hundred pages
- potential application to other text analysis tasks

Implemented components available online:
https://github.com/nik0spapp/icrawler

End of Presentation

Thank you! Any questions?

References


Dietrich Klakow M Wiegand, *Bootstrapping supervised machine-learning polarity classifiers with rule-based classification*, Proc. of the ECAI-Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA), 2009.


Nikolaos Pappas, Georgios Katsimpras, and Efstathios Stamatatos, *An agent-based focused crawling framework for topic- and genre-related Web document discovery*, 24th IEEE Int. Conf. on Tools with Artificial Intelligence (Athens, Greece), 2012.


Nikolaos Pappas, Georgios Katsimpras, and Efstathios Stamatatos, *Extracting informative textual parts from Web pages containing user-generated content*, 12th Int. Conf. on Knowledge Management and Knowledge Technologies (Graz, Austria), 2012.


Ellen Riloff and Janyce Wiebe, *Learning extraction patterns for subjective expressions*, Proc. of the 2003 Conf. on Empirical methods in natural language processing, EMNLP '03, 2003, pp. 105–112.


Theresa Wilson, Janyce Wiebe, and Paul Hoffmann, *Recognizing contextual polarity in phrase-level sentiment analysis*, Proc. of the Conf. on Human Language Technology and Empirical Methods in Natural