Voice of the Customers: Mining Online Customer Reviews for Product Feature-Based Ranking

Kunpeng Zhang, Ramanathan Narayanan, Alok Choudhary
Dept. of Electrical Engineering and Computer Science
Center for Ultra-Scale Computing and Security
Northwestern University
kzh980@eecs.northwestern.edu
ran310@eecs.northwestern.edu
choudhar@eecs.northwestern.eduz

3rd Workshop on Online Social Networks
Boston, MA, June 2010

1

Customer Reviews

- More consumers are shopping online than ever before

- Online retailers allow consumers to add reviews of products purchased

- Customer reviews are more unbiased, honest than product descriptions provided by sellers

Customers Shop Online and Read Reviews

★★★★★ **Canon gets it right with this mid-range SD series camera. Small and versatile**

By Anjana Nigam

Last year, Canon SD870IS camera was a winner in the SD series and the SD880IS manages to offer even more at a lower starting price (than that of SD870IS last year). *GENERAL POSITIVE COMPARATIVE*

A BIGGER CCD - 1/2.3" up from the previous 1/2.5" in the SD870 which allows for increase in MPs to 10MP (up from 8MP of SD870).

FEATURE: RESOLUTION , POSITIVE COMPARATIVE

- 4X Zoom (up from 3.8X on SD870) and a wide angle which really helps with shots at close quarters (in a car/bus, of a large group etc.

FEATURE: LENS , POSITIVE COMPARATIVE

- The huge 3 inch LCD screen is the same as the SD870 but slightly improved technology gives you brighter playback in low-light which makes it easier to see the shots in replay.

FEATURE: LCD , POSITIVE COMPARATIVE

It does not have a viewfinder but I have never missed the viewfinder on my SD870 since Canon made the screen anti-glare. I'd take a bigger screen any day over a 2.5 inch screen with a viewfinder, but if you like to use the viewfinder to compose your shots you might need to look at the SD990 IS and drop another 100 bucks.

FEATURE: LCD , NEGATIVE COMPARATIVE

4

Product Ranking Workflow

Online Data Crawling
_____
Information Extraction
_____
Sentence Splitter
_____
POS Tagging

Comparative Sentence Identification
_____
Comparative Sentence Refinement
_____
Subjective Sentence Identification

Graph Construction
_____
Product Ranking
_____

5

Product Features

Resolution

Lens

Memory

flash

Burst

compression

Focus

Battery
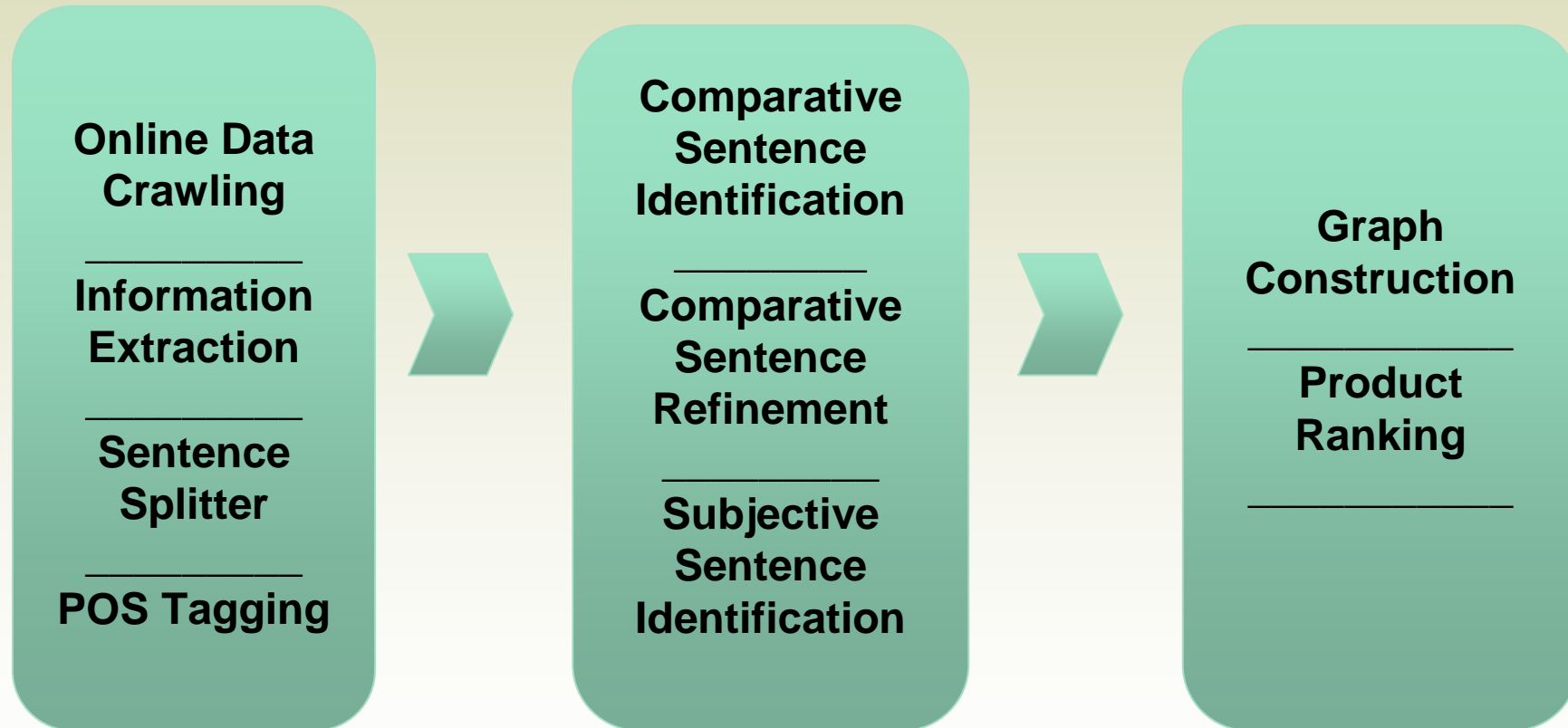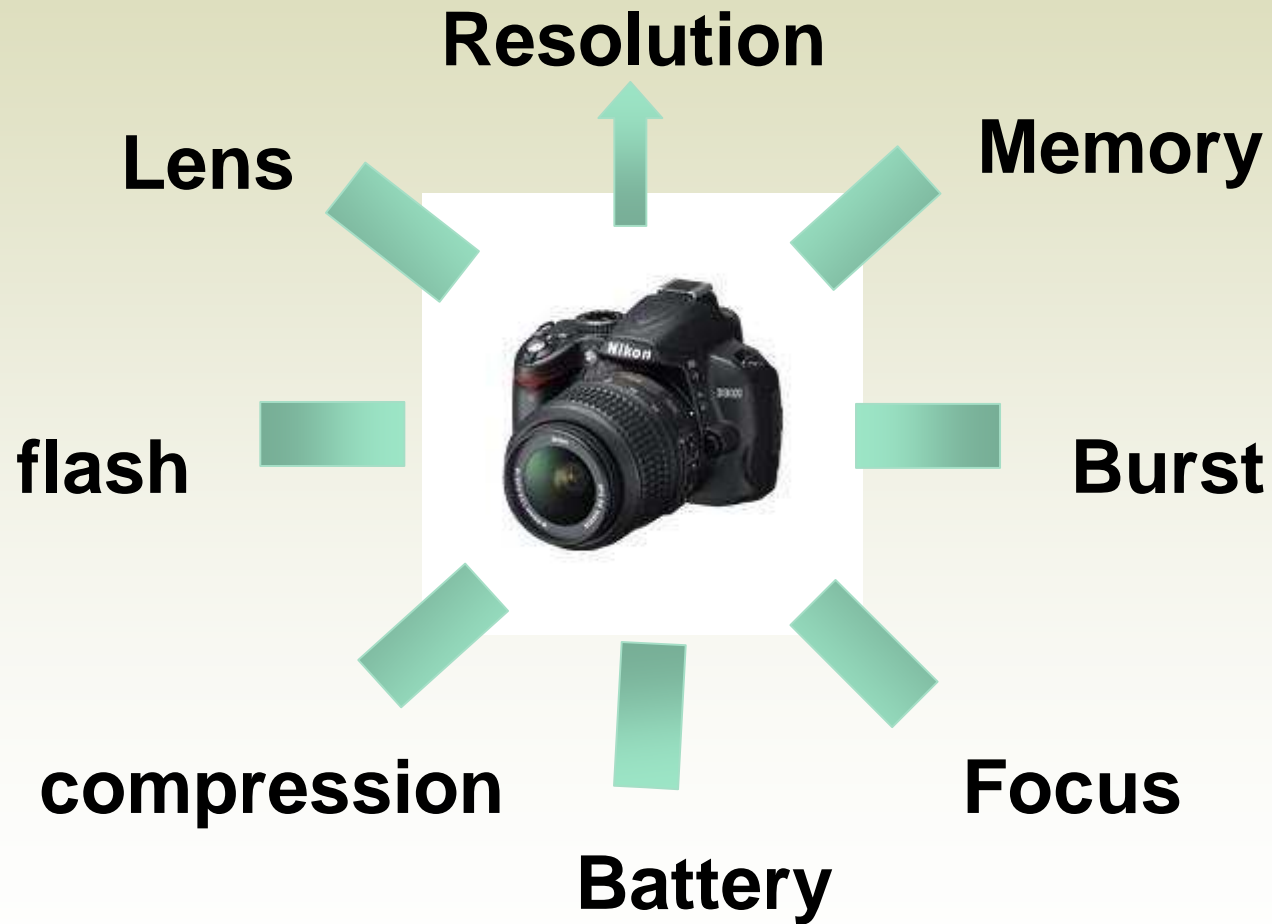
Identifying Feature Sentences

Table 1: Keywords Representing 10 Most Important Product Features for *Digital Camera* and *Television* Domains

Digital Camera	TV
resolution|pixel|megapixel	connection|input|output| component video|composite video|HDMI
lens|wide angle|normal range	adjustment|stretch|zoom|expand|compress
optical|zoom|optical zoom|digital zoom	film-mode|frame|theatrical|3:2|pull-down| motion compensation|CineMotion
memory|megabytes|MB	pip|picture-in-picture|dual-tuner|pop| picture-outside-picture|two-tuner
burst|continuous|shutter|recovery|motion|sport	resolution|1080p|1080i|720p
battery|batteries|power	screen|anti-glare|reflectivity|burn-in|shiny| screensaver|pixel-shift
focus|exposure|manual|iso	picture|image|picture quality|image quality
LCD|screen	sound|sound quality|speaker|stereo|audio
compression|compress|jpeg	size|height|width|depth|weight|inch
flash|light	remoter|remote|gear|universal

- Example: features from consumer reports

7

Sentence Labeling

Comparative sentence (CS)

A sentence which indirectly express an opinion by performing a comparison between two products

Rules to identify general CS

KW: 126 keywords (outperform, exceed…)

POS tags (JJR, RBR, JJS, RBS)

Predefined patterns (as <word> as…)

Refine CS

A sentence contains one or more than one different product names

8

Sentence Labeling

Subjective sentence (SS)

A sentence expressing directed praise or deprecation about a product

Rules to identify SS

If a sentence contains subjective words (positive, negative), it is classified to be a subjective sentence

Sentiment Identification

- ## Use the keyword strategy

  {MPQA[1] + our own words → 1974 positive words + 4605 negative words + 28 negation words}

  Positive subjective sentences (PS)

  Negative subjective sentences (NS)

  Positive comparative sentences (PC)

  Negative comparative sentences (NC)

- Positive Subjective(PS)
  - *This camera has great picture quality and conveniently priced.*
- Negative Subjective(NS)
  - *The picture quality of this camera is really bad.*
- Positive Comparative(PC)
  - *This camera has superior shutter speed when compared to the Nikon P40.*
- Negative Comparative(NC)
  - *This is the worst camera I have seen so far.*

[1]. *http://www.cs.pitt.edu/mpqa*

★★★★★ **Canon gets it right with this mid-range SD series camera. Small and versatile**

By Anjana Nigam

Last year, Canon SD870IS camera was a winner in the SD series and the SD880IS manages to offer even more at a lower starting price (than that of SD870IS last year). *GENERAL POSITIVE COMPARATIVE*

A BIGGER CCD - 1/2.3" up from the previous 1/2.5" in the SD870 which allows for increase in MPs to 10MP (up from 8MP of SD870).

FEATURE: RESOLUTION , POSITIVE COMPARATIVE

- 4X Zoom (up from 3.8X on SD870) and a wide angle which really helps with shots at close quarters (in a car/bus, of a large group etc.

FEATURE: LENS , POSITIVE COMPARATIVE

- The huge 3 inch LCD screen is the same as the SD870 but slightly improved technology gives you brighter playback in low-light which makes it easier to see the shots in replay.

FEATURE: LCD , POSITIVE COMPARATIVE

It does not have a viewfinder but I have never missed the viewfinder on my SD870 since Canon made the screen anti-glare. I'd take a bigger screen any day over a 2.5 inch screen with a viewfinder, but if you like to use the viewfinder to compose your shots you might need to look at the SD990 IS and drop another 100 bucks.

FEATURE: LCD , NEGATIVE COMPARATIVE

11

Constructing the Product Graph

- G = { *V, E* }
- *V* is the set of vertices. Each vertex represents a product
- *E* is the set of edges. The edge weight represents the comparative relationship between products
- Weight of a node is determined by the number of positive/negative subjective sentences (PS/NS)

  ex. *A* has excellent picture quality

- Weight of an edge is determined by the number of positive and negative comparative sentences

  ex. *B* is better than *C*. *B* is worse than *C*

12
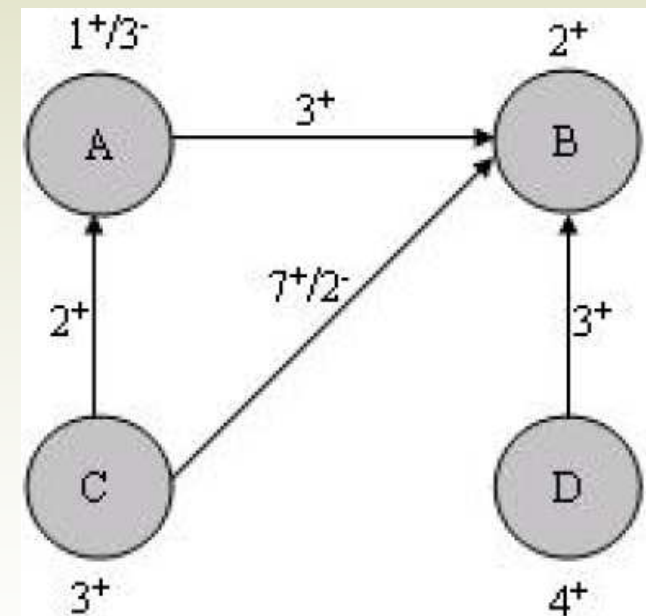
Ranking Products (pRank Algorithm)

For each graph (G_f) related to feature f, we could evaluate the relative importance of each product by using the pRank algorithm
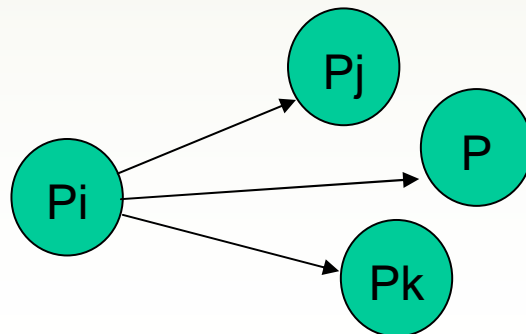
$$\mathbf{pRank}(P) = [(1 - d) + d * \sum_{i=1}^{n} \mathbf{1}_{\{P_i,P\}} * pRank(P_i) * C_e(P_i)] * C_v(P),\text{ where}$$

- pRank(P) is the product ranking of product P;

- pRank(P_i) is the product ranking of product P_i and n is the number of incoming links on product P;

- $\mathbf{1}_{\{P_i,P\}}$ is an indicator function, s.t.

$$\mathbf{1}_{\{P_i,P\}} = \begin{cases} 1 & \text{if there is a link from } P_i \text{ to } P \\ 0 & \text{otherwise} \end{cases}$$

- $C_e(P_i) = \frac{W_e(P_i,P)}{\sum_{j=1}^{m} W_e(P_i,P_j)}$, where m is the number of outbound links on product P_i, P_j are the nodes pointed to from P_i and $W_e(P_i, P_j)$ is the weight of the edge (P_i, P_j). It is the edge weight contributor to the ranking of product P;

- $C_v(P) = \frac{W_v(P,P)}{\sum_{t=1}^{n} W_v(P_t,P_t)}$. It is the node weight contributor to the ranking of product P.

13

Ranking Products (Example)
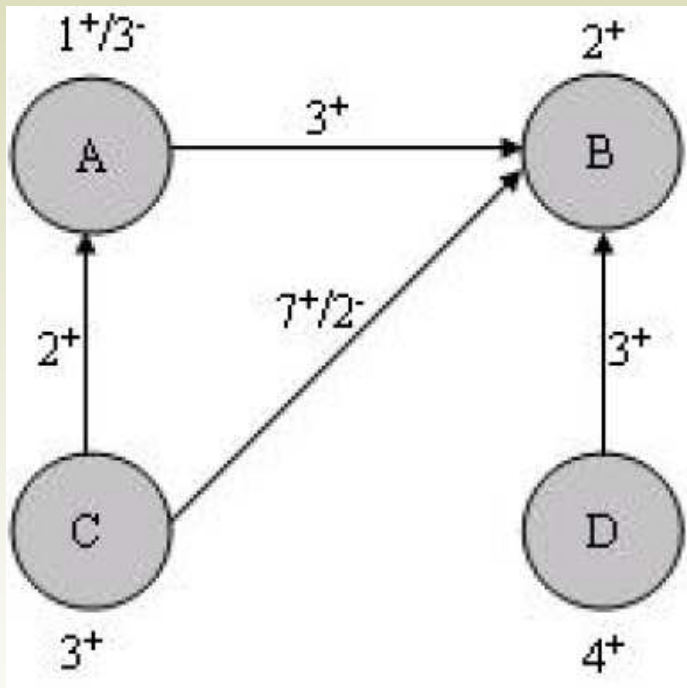
Rank	Product	Score
1	B	0.8
2	D	0.07
3	C	0.05
4	A	0.04

Example:

PS(A)=1, PS(B)=2, PS(C)=3, PS(D)=4,

NS(A)=3,

PC(B,A)=3, PC(B,C)=7, PC(B,D)=3, PC(A,C)=2,

NC(B,C)=2

14

Experiments

- Data

  - Amazon.com (Digital camera)

Table 3: Breakdown of Subjective/Comparative Sentences(Digital Camera)

Feature/Overall	No. of Sentences	No. of Subjective Sentences		No. of Comparative Sentences	
		Positive	Negative	Positive	Negative
Flash	48378	10045	8202	1358	514
Battery	42461	4838	6439	1030	533
Focus	42393	7306	7241	1389	720
Lens	36371	4678	5313	1055	437
Optical	28658	3771	3196	842	338
Lcd	25874	4357	3587	755	216
Resolution	14992	1768	1647	579	227
Burst	14362	2925	2726	523	189
Memory	10794	1225	1652	365	143
Compression	1780	225	236	78	29
Digital Camera	1469940	71565	97349	16246	10890

Precision and recall to identify feature sentences: *0.853* and *0.807*

15

Experiments

- ## Data

  - ### Amazon.com (TV)

Table 4: Breakdown of Subjective/Comparative Sentences(TV)

Feature/Overall	No. of Sentences	No. of Subjective Sentences		No. of Comparative Sentences	
		Positive	Negative	Positive	Negative
Sound	13877	1599	1933	456	303
Screen	9021	1374	1457	501	344
Size	7214	492	516	342	214
Connection	6299	465	641	239	163
Resolution	6155	286	306	418	256
Picture Quality	4987	2847	1750	201	65
Remoter	4554	619	715	175	117
Adjustment	1704	170	215	74	48
PIP	1205	139	175	49	43
Film-Mode	1022	167	158	53	23
TV	460610	17843	28510	10224	9162

16

Experiments

- Results for Camera and TV

Definition *Relative Feature Fraction:* $RFF_f = \frac{N_f}{\sum_f N_f} *100\%$, *where* N_f *is the number of sentences labeled with feature f.*

Definition *Importance of Feature:* $IF_f = \frac{|X \cap Y_f|}{|X|} *100$, *where* $X = \{top\ 10\%\ of$ *overall ranked products*$\}$, *and* $Y_f = \{top\ 10\%\ of\ products\ according\ to\ feature\ f\}$.

Digital Camera	RFF_f	TV	RFF_f
Flash	18.18%	Sound	24.76%
Battery	15.96%	Screen	16.10%
Focus	15.93%	Size	12.87%
Lens	13.67%	Connection	11.24%
Optical	10.77%	Resolution	10.98%
LCD	9.72%	Picture Quality	8.90%
Resolution	5.63%	Remoter	8.13%
Burst	5.40%	Adjustment	3.04%
Memory	4.06%	PIP	2.15%
Compression	0.67%	Film-Mode	1.82%

Digital Camera Features	IF_f	TV Features	IF_f
Lens	79.9	Size	78.7
Resolution	79.8	Film-Mode	72.3
Optical	77.5	Picture Quality	70.7
Focus	76.3	Connection	69.1
Memory	76.3	PIP	69.1
Burst	75.2	Sound	67.5
Lcd	74.1	Remoter	67.5
Flash	72.9	Adjustment	64.3
Battery	71.6	Screen	61.2
Compression	68.4	Resolution	61.2

17

- ## Top 10 Products for Each Feature and Overall

Lens	Resolution	Optical	Focus
PENTAX K100D SUPER	CANON POWERSHOT SD200	CASIO EXILIM PRO	CASIO EXILIM PRO
KODAK EASYSHARE ONE	NIKON COOLPIX S4	KODAK EASYSHARE ONE	KODAK EASYSHARE ONE
CANON POWERSHOT SD500	NIKON COOLPIX L3	CANON POWERSHOT SD550	PANASONIC DMC-FX37A 10.1MP
CANON POWERSHOT SD990IS	NIKON COOLPIX S600	CANON POWERSHOT SD990IS	CANON POWERSHOT S60
NIKON COOLPIX L3	KODAK EASYSHARE C653	NIKON COOLPIX 3200	CANON POWERSHOT SD990IS
NIKON COOLPIX 8400	NIKON COOLPIX P90	NIKON COOLPIX L3	NIKON COOLPIX 3200
NIKON COOLPIX L1	NIKON COOLPIX L4	NIKON COOLPIX S500	PENTAX OPTIO P70
KODAK EASYSHARE C653	SONY ALPHA A700K	HP PHOTOSMART M537	NIKON COOLPIX L3
NIKON COOLPIX P6000	NIKON COOLPIX P50	NIKON COOLPIX P50	NIKON COOLPIX 8400
NIKON COOLPIX L4	PENTAX OPTIO 60	NIKON COOLPIX 7600	NIKON COOLPIX L1

Memory	Burst	LCD	Flash
KODAK EASYSHARE ONE	CASIO EXILIM PRO	KODAK EASYSHARE ONE	KODAK EASYSHARE ONE
CANON POWERSHOT S80	CANON POWERSHOT SD500	PANASONIC DMC-FX37A 10.1MP	CASIO EXILIM PRO
PANASONIC DMC-FX37A 10.1MP	PANASONIC DMC-FX37A 10.1MP	CANON POWERSHOT S80	PANASONIC DMC-FX37A 10.1MP
PENTAX OPTIO A10	CANON POWERSHOT S80	PENTAX OPTIO A10	NIKON COOLPIX S200
CANON POWERSHOT SD990IS	CANON POWERSHOT S60	CANON POWERSHOT SD990IS	SONY DSCP150 7MP
NIKON COOLPIX 3200	CANON POWERSHOT SD990IS	NIKON COOLPIX 3200	CANON EOS 1D
OLYMPUS SP-550UZ 7.1MP	NIKON COOLPIX 3200	CANON POWERSHOT S400	NIKON COOLPIX 8400
NIKON COOLPIX 4300	NIKON COOLPIX 995	CANON POWERSHOT G3	NIKON COOLPIX L1
CANON POWERSHOT S410	CANON POWERSHOT S100	SONY DSCP150 7MP	NIKON COOLPIX P90
NIKON COOLPIX 8400	NIKON COOLPIX L3	NIKON COOLPIX S500	SONY ALPHA A700K

Battery	Compression	Overall Quality	
CANON POWERSHOT SD990IS	CANON POWERSHOT A620	NIKON COOLPIX L1	
NIKON COOLPIX S4	CANON POWERSHOT SD300	PANASONIC DMC-FX37A 10.1MP	
NIKON COOLPIX S500	KODAK EASYSHARE ZD710	CANON POWERSHOT A990IS	
NIKON COOLPIX L3	CANON POWERSHOT S100	NIKON COOLPIX P90	
KODAK EASYSHARE C653	NIKON COOLPIX 8700	HP PHOTOSMART M537	
HP PHOTOSMART M537	NIKON COOLPIX 4300	CANON POWERSHOT A70	
NIKON COOLPIX P90	NIKON COOLPIX L3	NIKON COOLPIX 8800	
NIKON COOLPIX P6000	NIKON COOLPIX P50	FUJIFILM FINEPIX A330	
NIKON COOLPIX L4	CANON POWERSHOT S230	KODAK EASYSHARE C653	

18

- # Comparing to Expert Opinion

  - ## Digital Camera and TV

Price Range	# of Products	10%	# of Top Rated Products(Expert Ranking)
<100	159	-	0
100~200	171	9/17	17
200~300	98	7/10	10
300~400	51	2/4	4
400~500	25	2/3	6
500~700	28	2/3	2
700~1000	24	-	0
>1000	29	2/3	2
Average Probability of Overlap		62.2%	

Price Range	# of Products	10%	20%	# of Top Rated Products(Expert Ranking)
<300	72	-	-	1
300~400	46	-	-	1
400~500	38	-	-	2
500~600	27	2/3	3/4	4
600~700	25	1/3	4/5	5
700~800	21	1/2	3/4	6
800~1000	47	3/5	5/9	9
1000~1500	62	2/6	8/12	18
1500~1800	18	1/2	2/4	15
1800~3000	24	1/2	1/4	6
>3000	8	-	-	5
Average Probability of Overlap		50%	62.3%	

- Comparing pRank to ARS (average rating score), Salesrank for Camera

Comparison of ranking overalpping with expert ranking to two baselines
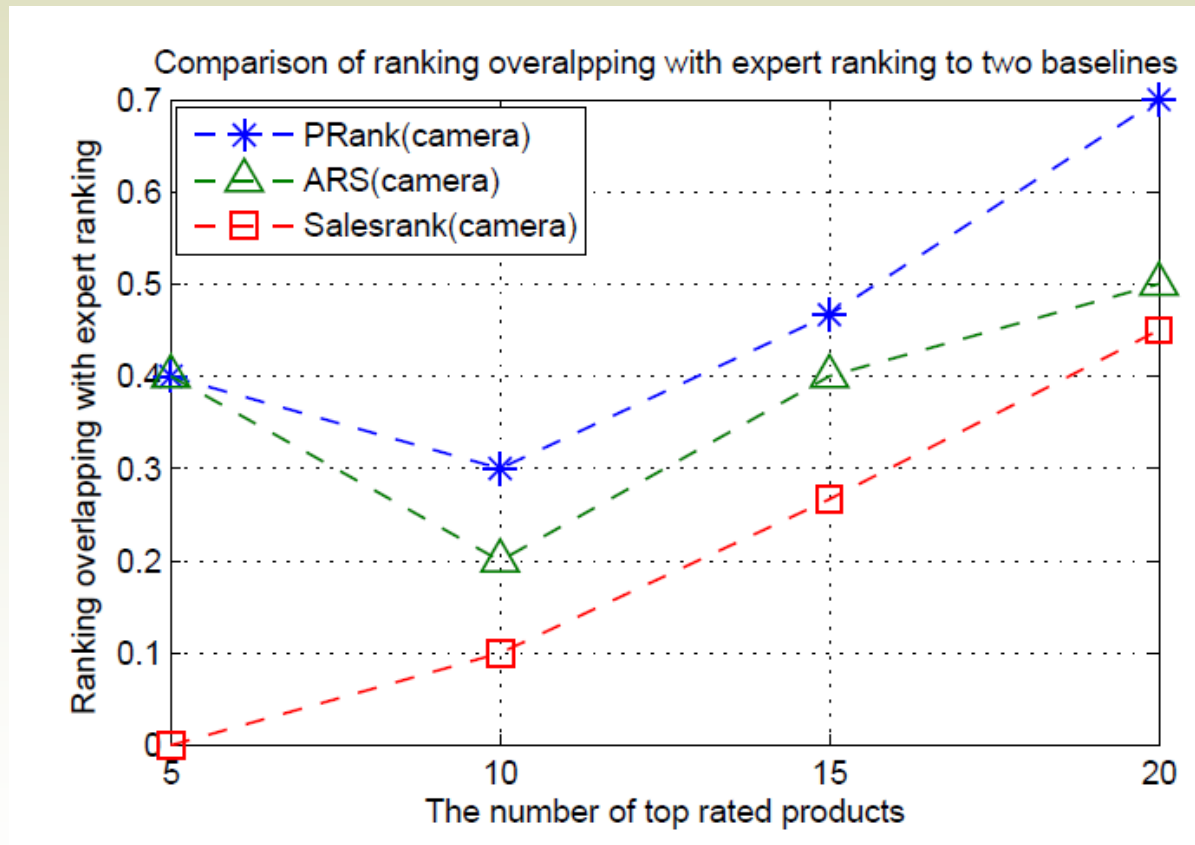
- **Comparing pRank to ARS (average rating score), Salesrank for TV**
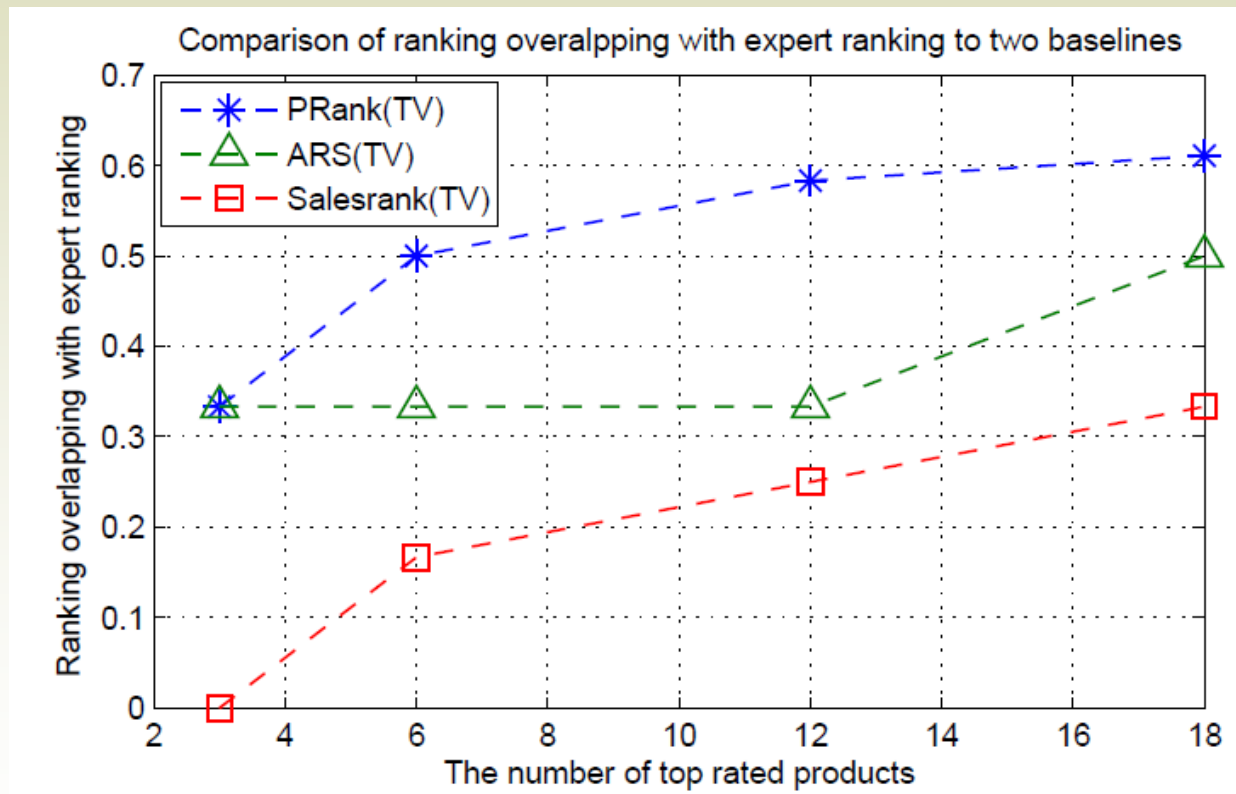
Related Work
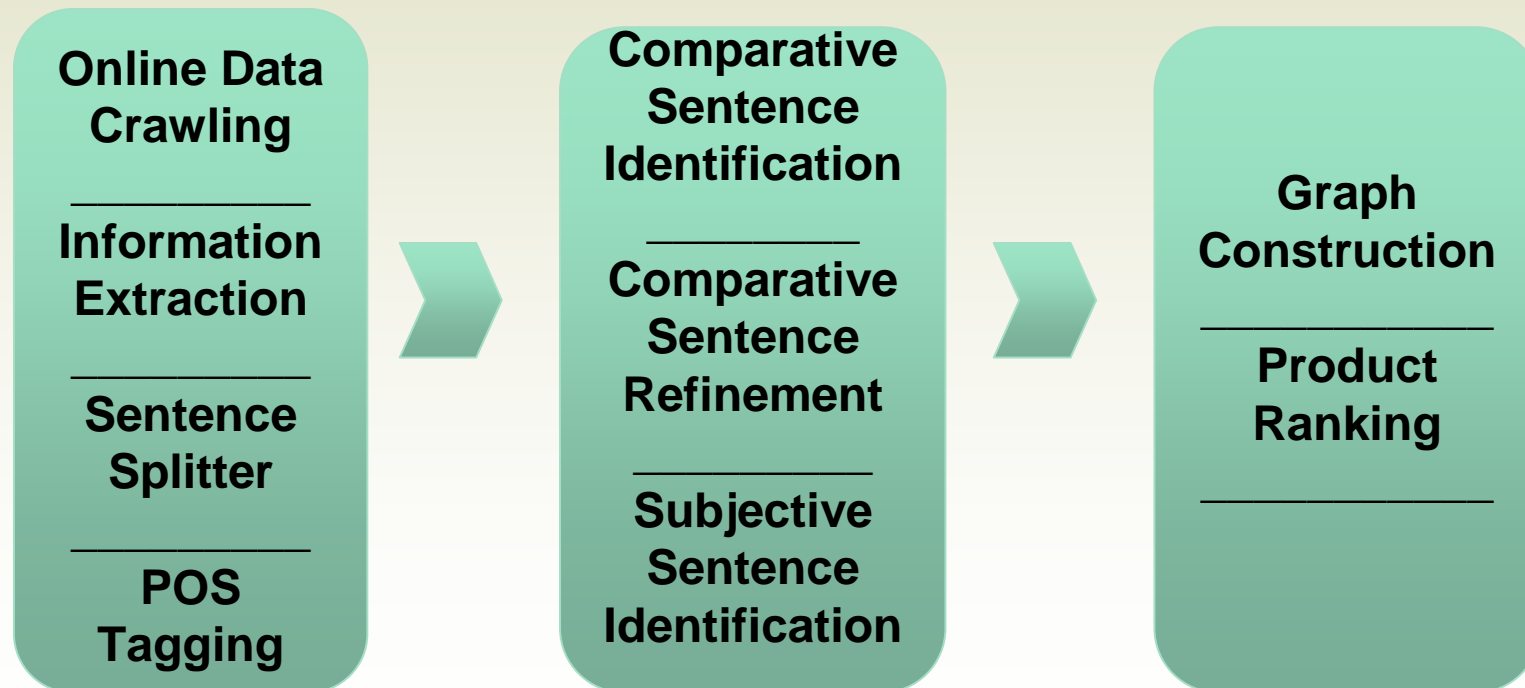
1. Sentiment analysis *[B. Liu, 2010; B. Pang, 2002]*

2. Extracting product features *[M. Hu, 2004; A. Popescu, 2005]*

3. Review summarization *[M. Hu, 2004, 2006]*

Differences from our work:

- Keyword matching strategy to identify and tag product features in sentences
- Different strategies to assign sentiment orientation to sentences
- Using our ranking algorithm on the product graph to rank products

Summary

Scalable technique to mine millions of online customer reviews to rank products

Online Data Crawling
—————————
Information Extraction
—————————
Sentence Splitter
—————————
POS Tagging

Comparative Sentence Identification
—————————
Comparative Sentence Refinement
—————————
Subjective Sentence Identification

Graph Construction
—————————
Product Ranking
—————————

Thank You

**Dept. of Electrical Engineering and Computer Science
Center for Ultra-Scale Computing and Security
Northwestern University**

**3rd Workshop on Online Social Networks
Boston, MA, June 2010**

Thank You

**Dept. of Electrical Engineering and Computer Science
Center for Ultra-Scale Computing and Security
Northwestern University**

**Workshop on Online Social Networks
Boston, MA, June 2010**