Incremental Cosine Computations for Search and Exploration of Tag Spaces

Raymond Vermaas, Damir Vandic, Flavius Frasincar

vandic@ese.eur.nl
http://damirvandic.com

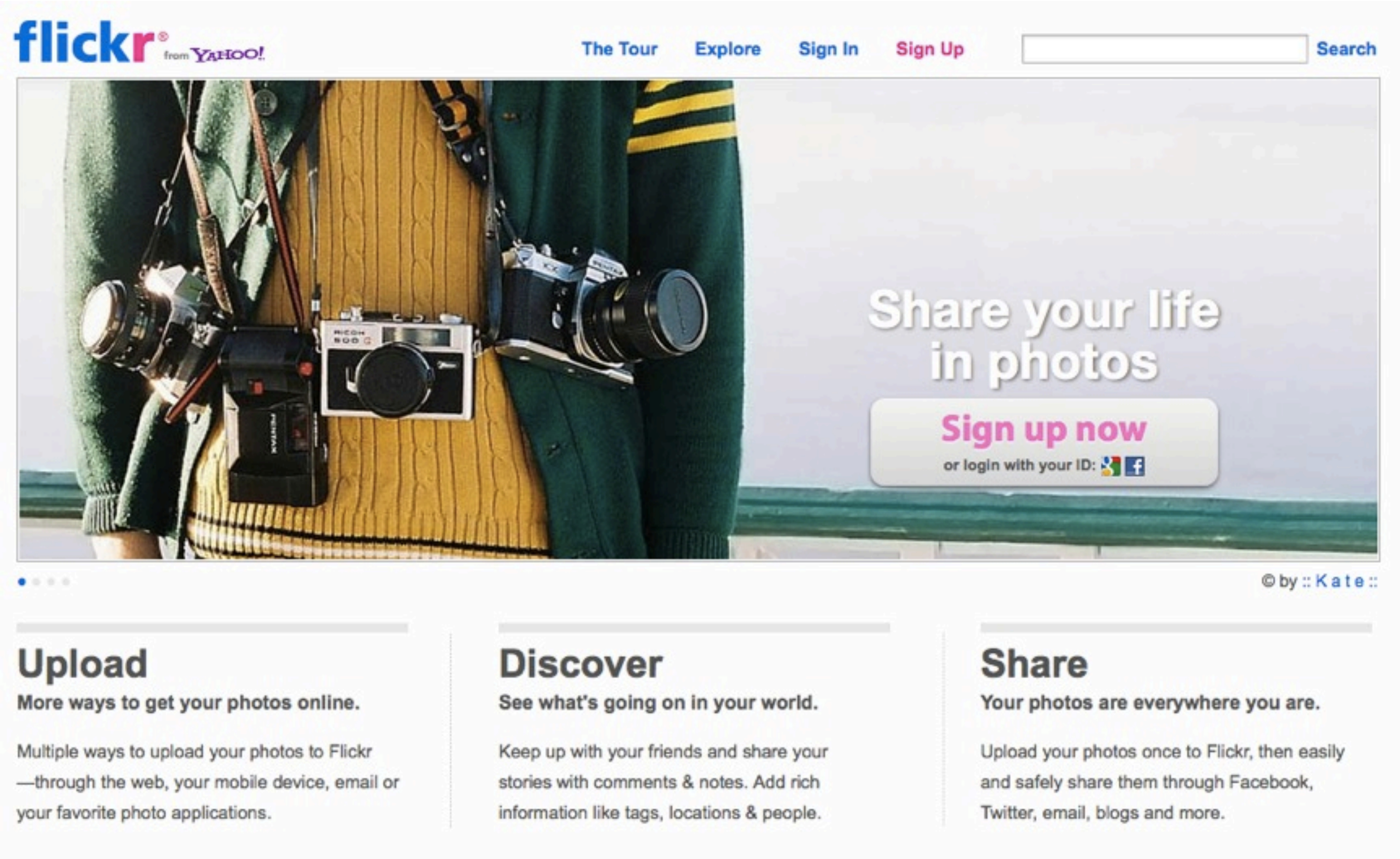
ERASMUS UNIVERSITEIT ROTTERDAM

Agenda

- Introduction and problem statement
- Two approaches
  1. the incremental recalculation approach
  2. the delta cosine approach
- Evaluation
- Discussion

Introduction

Introduction



Upload

More ways to get your photos online.

Multiple ways to upload your photos to Flickr —through the web, your mobile device, email or your favorite photo applications.

Discover

See what's going on in your world.

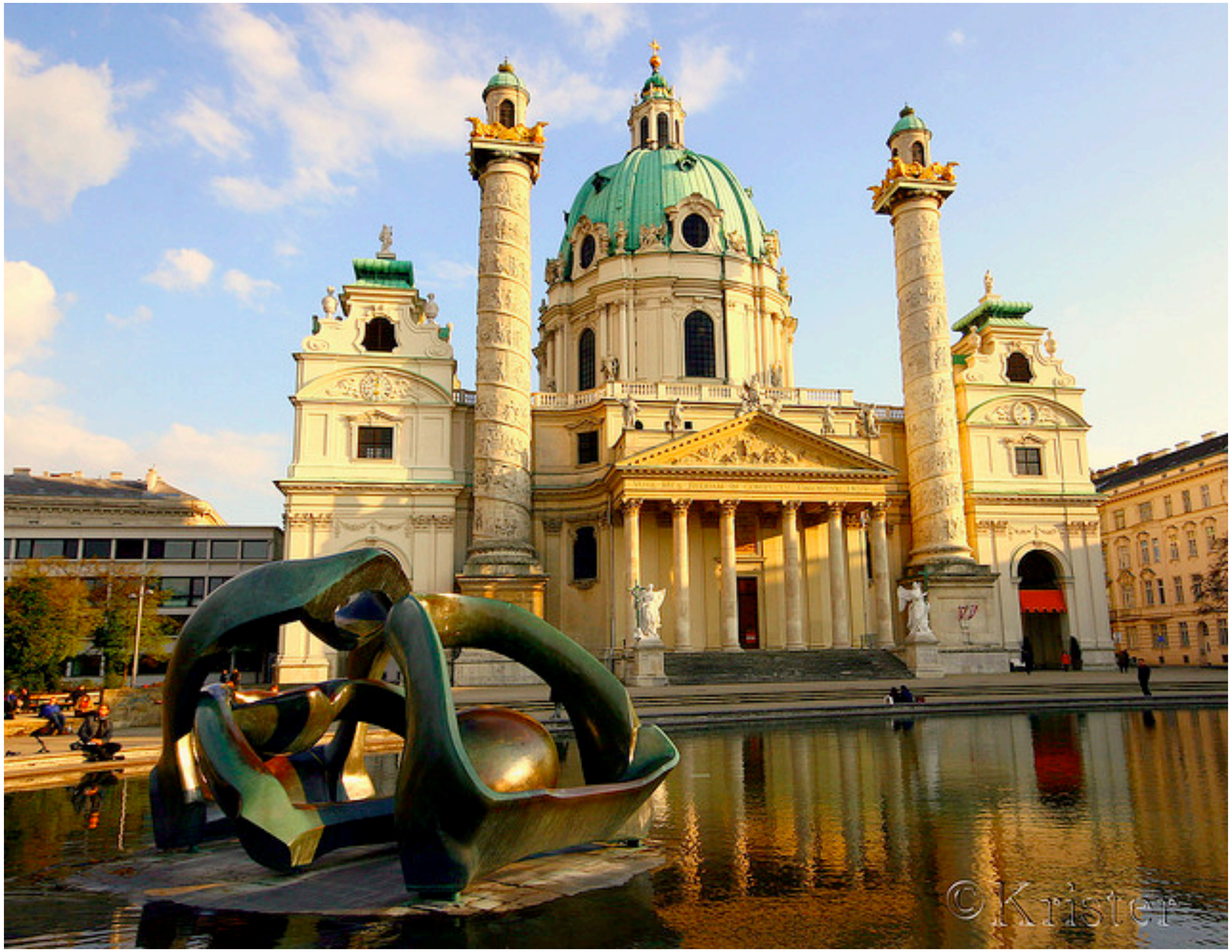
Keep up with your friends and share your stories with comments & notes. Add rich information like tags, locations & people.

Share

Your photos are everywhere you are.

Upload your photos once to Flickr, then easily and safely share them through Facebook, Twitter, email, blogs and more.

Introduction



- Canon
- 30D
- Sigma
- 10-20
- Vienna
- St. Charles's Church
- Karlskirche

Source: http://www.flickr.com/photos/krister462/2544707032/

Introduction

- Improving browsing and searching in tag spaces
  - clustering syntactic tag variations
  - clustering semantically related tags

Introduction

- Improving browsing and searching in tag spaces
  - clustering syntactic tag variations
  - clustering semantically related tags
- Employed similarity measure:
  - cosine similarity on tag co-occurrence vectors

Introduction

$$\cos(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{||\mathbf{a}|| \times ||\mathbf{b}||}$$

Introduction

$$\cos(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{||\mathbf{a}|| \times ||\mathbf{b}||}$$

	vienna	karlskirche	city	wiener schnitzel	rotterdam
vienna	-	2	3	2	0
karlskirche	2	-	0	0	0
city	3	0	-	0	4
wiener schnitzel	2	0	0	-	0
rotterdam	0	0	4	0	-

Problem

- Quadratic growth in number of cosines to compute

Problem

- Quadratic growth in number of cosines to compute
- Scalability problem when adding new images
  - new tags are introduced
  - co-occurrence update of existing tags

Problem

- How can we solve this?
  - an approach to ‘incrementally’ compute the cosines when new pictures are added

Problem

- How can we solve this?
    - an approach to ‘incrementally’ compute the cosines when new pictures are added
- Two approaches
    1. Incremental recalculation approach
    2. Delta cosine approach

Incremental Recalculation: co-occurrence update

Incremental Recalculation: co-occurrence update

tag	**1**	**2**	**3**	**4**	**5**	**6**
1	-	2	1	5	2	0
2	2	-	7	1	1	0
3	1	7	-	2	0	2
4	5	1	2	-	1	0
5	2	1	0	1	-	6
6	0	0	2	0	6	-

Incremental Recalculation: co-occurrence update

tag	**1**	**2**	**3**	**4**	**5**	**6**
1	-	2	1	5	2	0
2	2	-	7	1	1	0
3	1	7	-	2	0	2
4	5	1	2	-	1	0
5	2	1	0	1	-	6
6	0	0	2	0	6	-

Incremental Recalculation: co-occurrence update

tag	1	2	3	4	5	6
1	-	2	1	5	2	0
2	2	-	7	1	1	0
3	1	7	-	2	0	2
4	5	1	2	-	1	0
5	2	1	0	1	-	6
6	0	0	2	0	6	-

tag	1	2	3	4	5	6
1	-	2	1	5	2	0
2	2	-	7	1	1	0
3	1	7	-	3	0	2
4	5	1	3	-	1	0
5	2	1	0	1	-	6
6	0	0	2	0	6	-

Incremental Recalculation: co-occurrence update

tag	1	2	3	4	5	6
1	-	2	1	5	2	0
2	2	-	7	1	1	0
3	1	7	-	2	0	2
4	5	1	2	-	1	0
5	2	1	0	1	-	6
6	0	0	2	0	6	-

tag	1	2	3	4	5	6
1	-	2	1	5	2	0
2	2	-	7	1	1	0
3	1	7	-	3	0	2
4	5	1	3	-	1	0
5	2	1	0	1	-	6
6	0	0	2	0	6	-

Incremental Recalculation: co-occurrence update

tag	1	2	3	4	5	6
1	-	2	1	5	2	0
2	2	-	7	1	1	0
3	1	7	-	2	0	2
4	5	1	2	-	1	0
5	2	1	0	1	-	6
6	0	0	2	0	6	-

tag	1	2	3	4	5	6
1	-	2	1	5	2	0
2	2	-	7	1	1	0
3	1	7	-	3	0	2
4	5	1	3	-	1	0
5	2	1	0	1	-	6
6	0	0	2	0	6	-

In total 9 cosine similarities to update

Incremental Recalculation: adding a new tag

Incremental Recalculation: adding a new tag

tag	1	2	3	4	5	6	7
1	-	2	1	5	2	0	1
2	2	-	7	1	1	0	0
3	1	7	-	3	0	2	6
4	5	1	3	-	1	0	1
5	2	1	0	1	-	6	0
6	0	0	2	0	6	-	0
7	1	0	6	1	0	0	-

Incremental Recalculation: adding a new tag

tag	**1**	**2**	**3**	**4**	**5**	**6**	**7**
1	-	2	1	5	2	0	1
2	2	-	7	1	1	0	0
3	1	7	-	3	0	2	6
4	5	1	3	-	1	0	1
5	2	1	0	1	-	6	0
6	0	0	2	0	6	-	0
7	1	0	6	1	0	0	-

Incremental Recalculation: adding a new tag

tag	1	2	3	4	5	6	7
1	-	2	1	5	2	0	1
2	2	-	7	1	1	0	0
3	1	7	-	3	0	2	6
4	5	1	3	-	1	0	1
5	2	1	0	1	-	6	0
6	0	0	2	0	6	-	0
7	1	0	6	1	0	0	-

tag	1	2	3	4	5	6	7
1	-	2	1	5	2	0	1
2	2	-	7	1	1	0	0
3	1	7	-	3	0	2	6
4	5	1	3	-	1	0	1
5	2	1	0	1	-	6	0
6	0	0	2	0	6	-	0
7	1	0	6	1	0	0	-

Incremental Recalculation: adding a new tag

tag	1	2	3	4	5	6	7
1	-	2	1	5	2	0	1
2	2	-	7	1	1	0	0
3	1	7	-	3	0	2	6
4	5	1	3	-	1	0	1
5	2	1	0	1	-	6	0
6	0	0	2	0	6	-	0
7	1	0	6	1	0	0	-

tag	1	2	3	4	5	6	7
1	-	2	1	5	2	0	1
2	2	-	7	1	1	0	0
3	1	7	-	3	0	2	6
4	5	1	3	-	1	0	1
5	2	1	0	1	-	6	0
6	0	0	2	0	6	-	0
7	1	0	6	1	0	0	-

Incremental Recalculation: adding a new tag

tag	1	2	3	4	5	6	7
1	-	2	1	5	2	0	1
2	2	-	7	1	1	0	0
3	1	7	-	3	0	2	6
4	5	1	3	-	1	0	1
5	2	1	0	1	-	6	0
6	0	0	2	0	6	-	0
7	1	0	6	1	0	0	-

tag	1	2	3	4	5	6	7
1	-	2	1	5	2	0	1
2	2	-	7	1	1	0	0
3	1	7	-	3	0	2	6
4	5	1	3	-	1	0	1
5	2	1	0	1	-	6	0
6	0	0	2	0	6	-	0
7	1	0	6	1	0	0	-

In total 12 cosine similarities to update

Delta cosine approach: co-occurrence update

Delta cosine approach: co-occurrence update

tag	1	2	3	4	5	6
1	-	2	1	5	2	0
2	2	-	7	1	1	0
3	1	7	-	2	0	2
4	5	1	2	-	1	0
5	2	1	0	1	-	6
6	0	0	2	0	6	-

tag	1	2	3	4	5	6
1	-	2	1	5	2	0
2	2	-	7	1	1	0
3	1	7	-	3	0	2
4	5	1	3	-	1	0
5	2	1	0	1	-	6
6	0	0	2	0	6	-

Delta cosine approach: co-occurrence update

tag	1	2	3	4	5	6
1	-	2	1	5	2	0
2	2	-	7	1	1	0
3	1	7	-	2	0	2
4	5	1	2	-	1	0
5	2	1	0	1	-	6
6	0	0	2	0	6	-

tag	1	2	3	4	5	6
1	-	2	1	5	2	0
2	2	-	7	1	1	0
3	1	7	-	3	0	2
4	5	1	3	-	1	0
5	2	1	0	1	-	6
6	0	0	2	0	6	-

$$(\Delta \mathbf{t}_a + \mathbf{t}_a) \cdot \mathbf{t}_b = \mathbf{t}_a \cdot \mathbf{t}_b + \sum_{i \in U_a} \Delta \mathbf{t}_{ai} \times \mathbf{t}_{bi}$$

Delta cosine approach: co-occurrence update

Delta cosine approach: co-occurrence update

tag	**1**	**2**	**3**	**4**	**5**	**6**
1	-	2	1	5	2	0
2	2	-	7	1	1	0
3	1	7	-	2	0	2
4	5	1	2	-	1	0
5	2	1	0	1	-	6
6	0	0	2	0	6	-

tag	**1**	**2**	**3**	**4**	**5**	**6**
1	-	2	1	5	2	0
2	2	-	9	1	1	0
3	1	9	-	3	0	2
4	5	1	3	-	1	0
5	2	1	0	1	-	6
6	0	0	2	0	6	-

Delta cosine approach: co-occurrence update

tag	1	2	3	4	5	6
1	-	2	1	5	2	0
2	2	-	7	1	1	0
3	1	7	-	2	0	2
4	5	1	2	-	1	0
5	2	1	0	1	-	6
6	0	0	2	0	6	-

tag	1	2	3	4	5	6
1	-	2	1	5	2	0
2	2	-	9	1	1	0
3	1	9	-	3	0	2
4	5	1	3	-	1	0
5	2	1	0	1	-	6
6	0	0	2	0	6	-

$$(\Delta\mathbf{t}_a + \mathbf{t}_a)\cdot(\Delta\mathbf{t}_b + \mathbf{t}_b) = \mathbf{t}_a \cdot \mathbf{t}_b + (\sum_{i\in U_a} \Delta\mathbf{t}_{ai} \times \mathbf{t}_{bi}) + (\sum_{i\in U_b} \mathbf{t}_{ai} \times \Delta\mathbf{t}_{bi})$$

$$+ (\sum_{i\in U_{ab}} \Delta\mathbf{t}_{ai} \times \Delta\mathbf{t}_{bi})$$

Delta cosine approach: adding a new tag

- For the new tag:
  - use regular dot product

Delta cosine approach: adding a new tag

- For the new tag:
  - use regular dot product
- The resulting changes to the existing tags:
  - use the previously proposed formula’s (where the ‘old’ value is set to be 0)

Delta cosine approach: Euclidean norm

- For changes to existing tags:

$$||\varDelta\mathbf{t}_a + \mathbf{t}_a|| = \sqrt{||\mathbf{t}_a||^2 + \sum_{i \in U_a} (\varDelta\mathbf{t}_{ai} + \mathbf{t}_{ai})^2 - \sum_{i \in U_a} \mathbf{t}_{ai}^2}$$

(both due to updates to co-occ. and newly added tags)

Delta cosine approach: Euclidean norm

- For changes to existing tags:

$$||\mathit{\Delta}\mathbf{t}_a + \mathbf{t}_a|| = \sqrt{||\mathbf{t}_a||^2 + \sum_{i \in U_a} (\mathit{\Delta}\mathbf{t}_{ai} + \mathbf{t}_{ai})^2 - \sum_{i \in U_a} \mathbf{t}_{ai}^2}$$

  (both due to updates to co-occ. and newly added tags)

- For newly added tags:

$$||\mathbf{t}_a|| = \sqrt{\sum_{i=0}^{n} \mathbf{t}_{ai}^2}$$

Evaluation

- Consider execution time
  - Recalculate everything included as baseline approach

Evaluation

- Consider execution time
  - Recalculate everything included as baseline approach
- Initial data set contains 50,000 pictures and 1,444 tags
  - 8 incremental data sets are used to simulate pictures flowing into the system

Evaluation

New pictures	2,500	5,000	12,500	25,000	37,000	50,000	62,500	75,000
New pictures (%)	5%	10%	25%	50%	75%	100%	125%	150%
New tags	1	22	183	712	1,408	2,193	2,890	3,682
Total nr. of tags	1,445	1,466	1,627	2,156	2,852	3,637	4,334	5,126
Updated co-occurrences	10,319	33,730	52,351	79,482	98,643	112,905	137,723	140,145
Updated co-occurrences (%)	1.0%	3.2%	5.0%	7.6%	9.5%	10.8%	13.2%	13.5%

Evaluation

New pictures	2,500	5,000	12,500	25,000	37,000	50,000	62,500	75,000
Time (s) complete recalculation	23	23	31	74	179	378	677	1,137
Time (s) incremental	16	17	24	58	144	304	551	922
Time (s) delta	1	1	8	39	120	288	538	919
Speed-up (delta vs complete)	23	23	3.9	1.9	1.49	1.31	1.26	1.24
Speed-up (delta vs incremental)	16	17	3	1.48	1.2	1.06	1.02	1.00

Any questions?