



Integrating Tags in a Semantic Content-based Recommender

Jae Yong Kim

Degemmis, Marco & Lops, Pasquale & Semeraro, Giovanni & Basile, Pierpaolo. (2008). Integrating tags in a semantic content-based recommender. RecSys'08: Proceedings of the 2008 ACM Conference on Recommender Systems. 163-170. 10.1145/1454008.1454036.

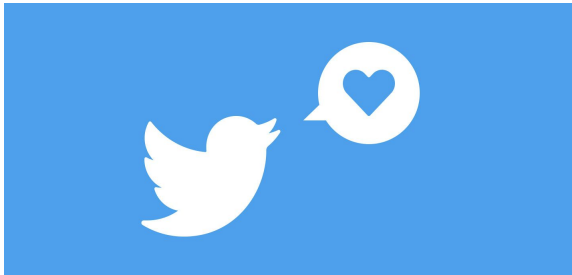
Contents



1. Introduction
2. ITR (ITem Recommender) System Process
 - a. Content Analyzer
 - b. User Profile Learner
 - c. Recommender
3. Content Analyzer - Semantic Indexing
4. User Profile Learner - Multivariate Poisson Model
5. Augmenting Tags to ITR
6. Experiment / Dataset
7. Results
8. Conclusion

Introduction

- User roles: Recipient → Participant
- folksonomy => folks + taxonomy == #Tags
 - Socially constructed classification schema
- Does the integration of tags cause an increase of the prediction accuracy in the process of recommending items to users?



Recommender System

1. Content Analyzer
 - a. Semantic Indexing
2. Profile Learner
 - a. Multivariate Poisson Model
3. Recommender
 - a. ITR (ITem Recommender)
+ UGC(User Generated Contents)

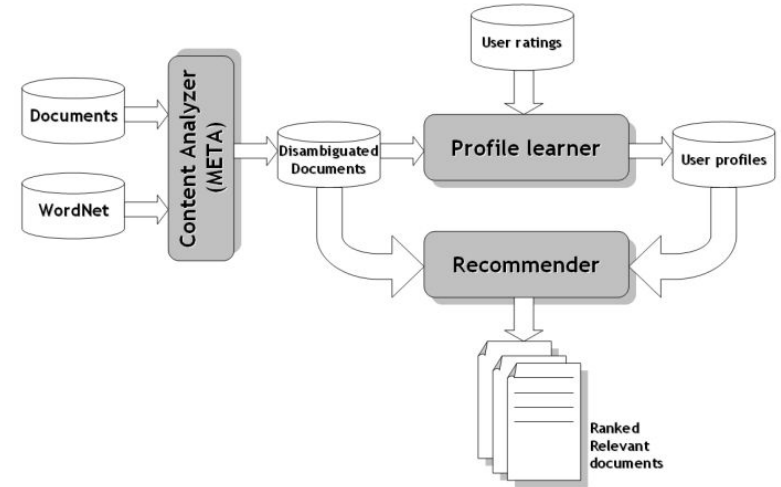


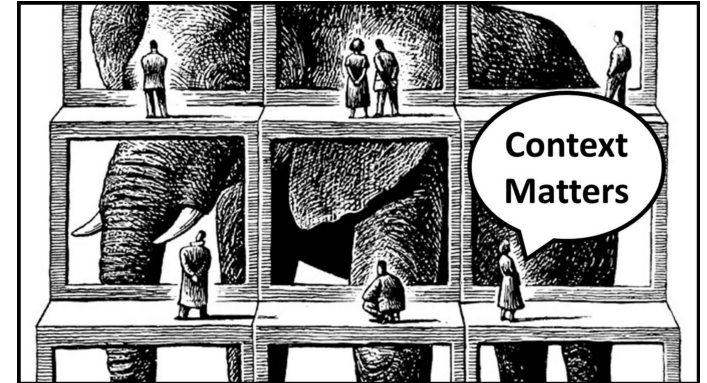
Figure 1: ITR architecture

Content Analyzer

- Documents: Textual description of items

Relevant *Concepts* surrounding the content (vs. keywords?)

- final output: Disambiguated document
- How?
 - repository for word senses \Rightarrow WORDNET 2.0
 - Word Sense Disambiguation (WSD)



WSD - JIGSAW

Determining which word is right in the situation

$$d = [w_1, w_2, \dots, w_h]$$

Semantic similarity = relatedness between word (A, B) --> Leacock-Chodorow measure

$$\text{sim}_{\text{LC}}(c_1, c_2) = -\log \frac{\text{len}(c_1, c_2)}{2 \times \max_{c \in \text{WordNet}} \text{depth}(c)}$$

$$\text{JIGSAW}(d) = \text{WORDNET synsets (Synonym-set)}X = [s_1, s_2, \dots, s_k] \quad (k \leq h)$$

BOS (Bag-of-Synsets)

** BOS: Synset-based vector space representation

Textual Slots : item property representation

$$d_n^s = \langle t_{n1}^s, t_{n2}^s, \dots, t_{nD_{ns}}^s \rangle \xrightarrow{\text{rep in vector space}} f_n^s = \langle w_{n1}^s, w_{n2}^s, \dots, w_{nD_{ns}}^s \rangle$$

s = index of slot

w = weight of synset

n = nth document in N-documents

(frequency of synset t_n)

t = set of all different synsets found in slot

Learning User Profile

Multivariate Bernouli vs Multinomial Model

Problems

1. Variation in length of documents
2. Rare categories (Not enough samples)

Let's use the Poisson distribution(model) for learning the bayes text classifier !

$$\begin{aligned} P(c|d_j) &= \frac{P(d_j|c)P(c)}{P(d_j|c)P(c) + P(d_j|\bar{c})P(\bar{c})} \\ &= \frac{\frac{P(d_j|c)}{P(d_j|\bar{c})}P(c)}{\frac{P(d_j|c)}{P(d_j|\bar{c})}P(c) + P(\bar{c})} \end{aligned} \quad (1)$$

If we set:

$$z_{jc} = \log \frac{P(d_j|c)}{P(d_j|\bar{c})} \quad (2)$$

then Eq. (1) can be rewritten as:

$$P(c|d_j) = \frac{e^{z_{jc}}P(c)}{e^{z_{jc}}P(c) + P(\bar{c})} \quad (3)$$

Multivariate Poisson Model

$$z_{jc} = \sum_{i=1}^{|V|} w_{ij} \cdot \log \frac{\lambda_{ic}}{\mu_{ic}}$$

V: Vocabulary size

w: frequency term of t in document d

$$z_{jc}^s = \sum_{i=1}^{|V|} w_{ij}^s \cdot \log \frac{\lambda_{ic}^s}{\mu_{ic}^s}$$

$$\lambda_{ic} = \frac{\text{\#occurrences for } t_i \text{ in the pos. training documents}}{\text{\#total tokens in the pos. training documents}},$$
$$\mu_{ic} = \frac{\text{\#occurrences for } t_i \text{ in the neg. training documents}}{\text{\#total tokens in the neg. training documents}}.$$

$$\lambda_{ic}^s = \frac{1}{|D_c|} \sum_{j=1}^{|D_c|} \hat{w}_{ij}^s \quad \mu_{ic}^s = \frac{1}{|D_{\bar{c}}|} \sum_{j=1}^{|D_{\bar{c}}|} \hat{w}_{ij}^s \quad s = 1, \dots, M \quad (10)$$

where D_c ($D_{\bar{c}}$) is the number of documents in class c (\bar{c}),

$$\hat{w}_{ij}^s = \frac{w_{ij}^s}{\alpha \cdot \text{avg}t f_j^s + (1 - \alpha) \cdot \text{avg}t f_j^s} \quad (11)$$

Training



- User has some discrete scale (MIN and MAX)
- positive training set if ratings $> (\text{MIN} + \text{MAX}) / 2$
- negative training set if ratings $< ((\text{MIN} + \text{MAX}) / 2)$
- compute a-posteriori classification scores $P(c+|d_j)$ and $P(c-|d_j)$, given new document d_j

Augmenting Recommender

ITR += static documents + dynamic user generated content(tags)

Tags => SocialTags(I), PersonalTags(U, I), PersonalTags(U)

1. WSD (JIGSAW): **Use static content as context instead of other tags
2. Profile learner : infers the profile as a binary text classifier
3. a-priori probabilities of profile_like and profile_dislike

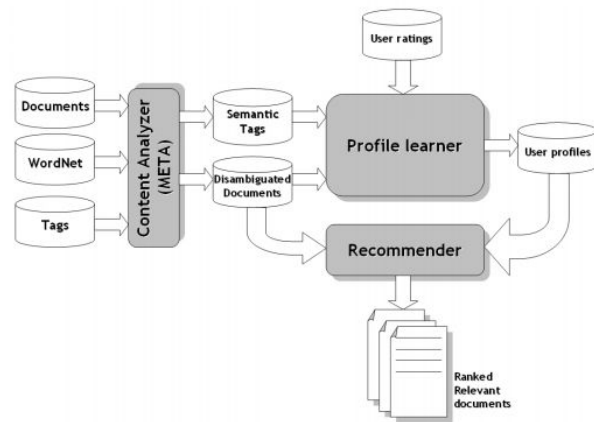


Figure 2: ITR 2.0 architecture

Experiment / Datasets

45 paintings chosen from the collection of the Vatican picture-gallery

- title, artist, description + tags and preference score on 5 points scale (1 = strongly dislike, 5= strongly like)

#1 only static

#2 only SemanticPersonal

#3 only SemanticSocial

#4 static + SemanticPersonal


#5 static + SemanticSocial

Accuracy: *Precision and Recall*

Precision(Pr): number of relevant selected items / number of selected items

Recall(Re) : number of relevant selected items / total number of relevant items available

27) Caravaggio - Deposition from the Cross



Painting Description

The Deposition, considered one of Caravaggio's greatest masterpieces, was commissioned by Girolamo Vintice for his family chapel in S. Maria in Vallicella (Chiesa Nuova) in Rome. In 1797 it was included in the group of works transferred to Paris at execution of the Treaty of Tolentino. After its return in 1817 it became part of Pisa VII's Pinacoteca. Caravaggio did not read portray the burial or the Deposition in the traditional way, inasmuch as Christ is not shown at the moment when he is laid in the tomb, but rather when, in the presence of the holy women, he is laid by Nicodemus and John on the Anonizing Stone, that is stone with which the sepulcher will be closed. Around the body of Christ are the Virgin, Mary Magdalene, John, Nicodemus and Mary of Cleophas, who raises her arms and eyes to heaven in a gesture of high-dramatic tension. Caravaggio, who arrived in Rome towards 1592-93, was the protagonist of a real artistic revolution as regards the way of treating subjects and the use of colour and light, and was certainly the most important forerunner of the 'realist' trend of seventeenth century painting.

Popular Tags: caravaggio (3) deposition (5) cross (4) christ (2) vamped (1) middle (1) unction (1) sepulchre (1) nicodemus (1) virgin (1)

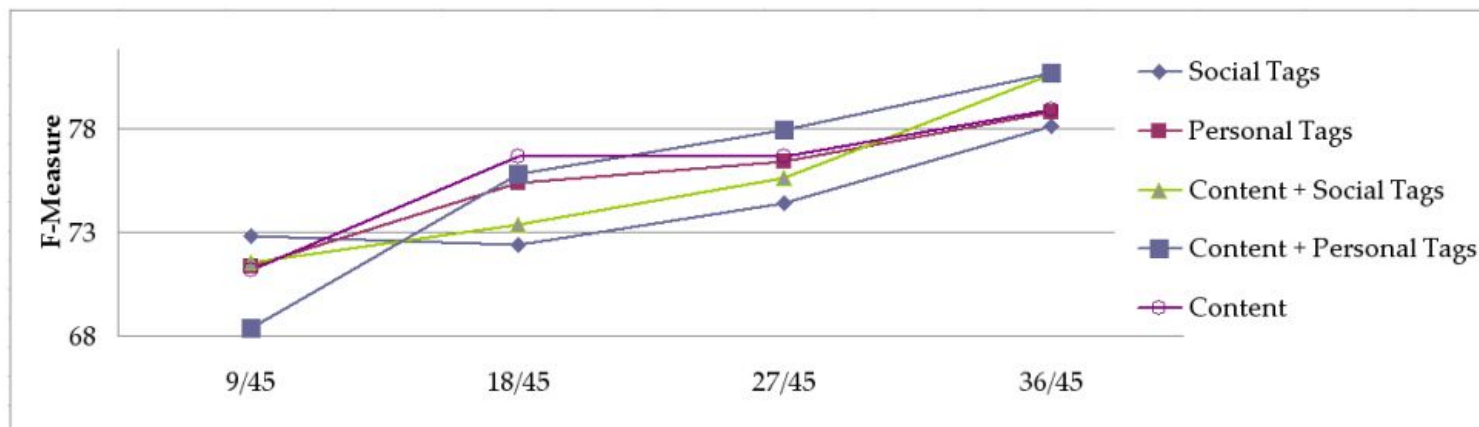
-Rate this painting and enter comma separated tags:
1 0 2 3 4 5 0

$$F_{\beta} = \frac{(1 + \beta^2) \cdot Pr \cdot Re}{\beta^2 \cdot Pr + Re}$$

Results

Table 2: Results of the K-fold Cross Validation

Type of content	Pr	Re	$F_{\beta=0.5}$
Static Content	75.86	94.27	78.94
SemanticPersonalTags(U,I)	75.96	92.65	78.80
SemanticSocialTags(I)	75.59	90.50	78.17
Static Content+SemanticPersonalTags(U,I)	78.04	93.60	80.72
Static Content+SemanticSocialTags(I)	78.01	93.19	80.64



Conclusion



Main contribution:

Multivariate Poisson model for naive Bayes text classification adapted to infer user profiles

- In the end, using tags along with static information is better than recommending through just keywords or static information itself!
- perform an analysis of what tags are used to build the folksonomies and how they affect the user profile generation
- More diverse users



Thank you!