

Content-Based Filtering Using **TFIDF**

YBIGTA 16기 박솔희





CONTENTS

01 Introduction 3p

02 TFIDF 4p

03 Content Based Filtering 5p



01 Introduction



02 TFIDF



03 Content Based
Filtering

01 Introduction

01

News

02

Clothing

03

Movies

04

Hotels

01 Introduction



**Model items
according to relevant attributes**



**Model user preferences
by attribute**

01 Introduction

Personalized Newspaper



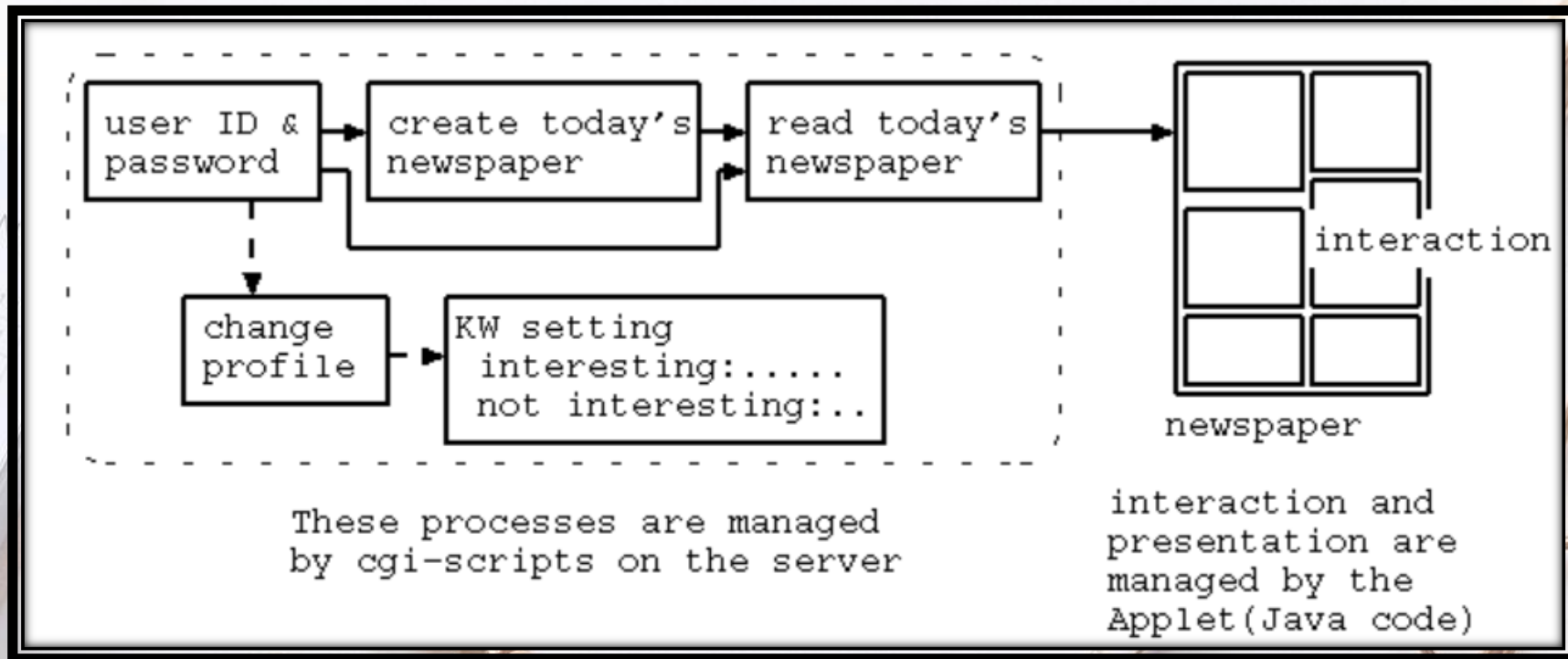
Amount



Position

01 Introduction

Reading a newspaper (User's view)



01 Introduction

Profile 제작

- User could build own profile
 - Infer profile from user actions
 - Infer profile from ratings
- Keyword에 대한 Likes와 dislikes을 count

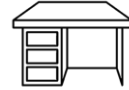
01 Introduction

etown's Ask Ida



Select

01



Choose

02



Decide

03



More

04

01 Introduction

etown's Ask Ida

The screenshot displays the etown.com website, which is titled 'THE Consumer Electronics Source'. The header includes a search bar, a date of July 17, 2001, and a navigation bar with links for 'Customer Service', 'Your Account', and 'Shopping Cart'. A banner below the header asks 'Should I buy a digital (DV format) camcorder?' with a 'click here' link.

The main content area is divided into two columns: 'News & Views' and 'Browse & Buy'.

News & Views

- This Just In**
 - Dynamo shoes generate electricity
 - Showtime phases in Dolby 5.1 sound
 - EMI to vend pay-per-download songs
 - Streaming A/V on your cell phone?
 - Stephen King plans serial e-book
 - Juno, Hughes team on Web by satellite
 - Will Sony Palm unit hit U.S. by Xmas?
- New Reviews/Products**
 - EXCLUSIVE:** Toshiba 480p-out DVD
 - Bush TV center has nice price
 - Vidikron projector: 'on eye-opener!'
 - Panamax nixes electrical spikes
 - JVC boombox has stylish looks
 - DSS security recommends AT&T phone
 - EXCLUSIVE:** Sharp DVD player
 - EXCLUSIVE:** Philips/TiVo recorder
 - EXCLUSIVE:** Outlaw 6.1 A/V receiver
- Features / Columns**
 - Unclear on the THX concept?
 - In the Mix: Did we ask for DVD-A?
 - Don't miss our updated FAQ for DTV!
 - Camcorder Corner: Fade to black
 - How to set up your subwoofer

Browse & Buy

- Home Theater**
 - VCR Satellite DVD Receivers
 - Surround Separates Speakers
 - Paired/Stereo Receivers
 - Accessories Etc.
- Television**
 - Direct View TV Rear Projection TV
 - Front Projection TV Flat Panel TV
 - High Definition Personal TV HDTV
- Camcorders**
 - Mini Format VHS Format
 - DV Format Digital Cameras
- Home Audio**
 - CD Player DVD/Blu-ray Recorders
 - Speakers Compact Systems Phono
 - Integrated Amps Stereo Separates
 - Accessories Multimedia Speakers
 - A/V Furniture
- Portable Tech**
 - Boomboxes Personal Cassette
 - Personal CD MP3/MP4 Palm PDA
 - Portable Radio Digital Cameras
 - MP3 Player Headphones
 - Zune Radio GPS
- Telecom**
 - cordless Phones Cell Phones
 - Home/Personal Wireless Pagers
 - Corded Telephones Fax Machines
 - Answering Machines
- Accessories**
 - Remote Controls Audio Cables
 - Video Cables A/V Furniture External
 - Cases/Bags Screen Protectors

Top 10 best selling items:

- Summer Sale**
 - Home brand products at discount prices
 - Great savings on every item!
 - Cambridge SoundWorks**
 - Model 88 Tabletop Radio
 - Was Price \$249.99
 - Sale \$149.99**
 - Sharp MM-877**
 - Model 88 Portable Recorder
 - Was Price \$299.99
 - \$169.70**
 - JVC GZ-OM60**
 - Was Price \$149.99
 - \$879.88**
 - JVC PND-018**
 - Was Price \$149.99
 - \$879.88**

01 Introduction



Select

01

etown's Ask Ida

Ida: Your Interactive Decision Assistant



Welcome!

I'm Ida, your interactive decision advisor. I can help you find the products that best suit your needs and preferences. I combine **etown.com**'s expertise on consumer electronics with state-of-the-art artificial-intelligence techniques from **Ask Jeeves**. I think you'll find shopping for electronics can be easy and fun!

Select the product category you are interested in:

Home Theater

- [DVD Players](#)
- [A/V Receivers](#)
- [VCRs](#)

Portable Tech

- [Digital cameras](#)
- [Boomboxes](#)
- [Handheld/Palm PC](#)

Home Audio

- [Compact Systems](#)
- [CD Players](#)

Telecom

- [Cordless Phones](#)

Camcorders

- [8mm, VHS, and DV](#)

01 Introduction

etown's Ask Ida



Choose

02



Choosing a digital camera

Picking out the right digital camera is a matter of finding out which features and benefits will work best for you. Answer a few questions for us, and we'll help you pick one that you'll like. There are 50 digital cameras to choose from so let's get started.

How are you planning on using the images that you'll shoot with your new digital camera? (Check all that apply.)

- ☐ **Post them to a Web site.** I need enough picture quality for my shots to look good on screen, though I don't need to print them.
- ☐ **Email them to friends and family.** I need a wide range of picture quality, some people may want to print the pictures I send.
- ☐ **Make prints out of them.** I need an upper-level camera with the best possible picture resolution, because at some point, there'll be hard copies.



Experience

Your answers to the next 2 questions will help me determine which digital cameras I should recommend to you.

In general, how experienced a photographer are you?

- ☐ **I'm a casual user.** I like to shoot pictures, and I don't like to fuss with technology. The simpler my camera is, the better I like it.
- ☒ **I know my way around a camera.** I'm interested in a digital camera with features that can make my work better and more enjoyable.
- ☐ **I'm an avid photographer.** I want an advanced digital camera that can keep up with my ideas and provide me with the most creative options.


01 Introduction

etown's Ask Ida

- **Olympus D450Z** [See etown.com Review](#) **Buy \$499**
Pros: It can store 18 pictures at its highest resolution, it has 1280 x 960 pixels resolution, it uses SmartMedia to store pictures, and it has an optical zoom lens.
- **Olympus D460 Zoom** **Buy \$499**
Pros: It can store 18 pictures at its highest resolution, it has 1280 x 960 pixels resolution, it uses SmartMedia to store pictures, and it has an optical zoom lens.
- **Fuji MX1200** [See etown.com Review](#) **Buy \$299**
Pros: It can store 23 pictures at its highest resolution, it has 1280 x 960 pixels resolution, and it uses SmartMedia to store pictures. **Cons:** It doesn't have an optical zoom lens.
- **Nikon 990** **Buy \$599**
Pros: It has 1600 x 1200 pixels resolution, it has CompactFlash storage media, and it has an optical zoom lens. **Cons:** It can store only 8 pictures at its highest resolution.

[Top of Page](#)

I can refine these recommendations if you tell me more about your needs. I suggest **Optical zoom** as the next question to consider, or you can select the topic you wish:

[Next question](#) 

03



Decide

01 Introduction

etown's Ask Ida



✓ **Nikon 800 - \$599.00 (msrp)**

Given the information you have provided, the Nikon 800 is one of my top recommendations. Click the link to go directly to a question that will explain the feature and help you decide if that feature makes sense for you!

Pros: Its advantages include:

- it has an LCD view screen.
- it has manual overrides.
- it has an optical viewfinder.
- it has 1600 x 1200 pixels resolution.
- it has a serial output connection.
- it has CompactFlash storage media.
- it has a video out connection.
- it has a built-in digital zoom.
- it has an optical zoom lens.

Cons: Possible disadvantages include:

- it can store only 8 pictures at its highest resolution.
- it doesn't have a USB connection.

04



More

01 Introduction

Entrée's Navigation Interface



01 Introduction

Challenges and Drawbacks

- Painting의 경우, 특징들을 listing하기가 어렵다
- 각각의 특징들이 Item에 적절하게 분배되어 있어야 한다
- 색다른 Connection을 찾기가 어렵다
- Substitutes을 찾기는 쉬워도, Complements를 찾기 어렵다



01 Introduction



02 TFIDF



03 Content Based
Filtering

02 TFIDF

TF * IDF

TF : Term Frequency

특정한 단어가 문서 내에 얼마나 자주 등장하는지를 나타내는 값

IDF : Inverse Document Frequency

여러 문서에서 등장한 단어의 가중치를 낮추는 역할

$$\text{idf}(t, D) = \log \frac{|D|}{|\{d \in D : t \in d\}|}$$

02 TFIDF

TFIDF의 역할

- 모든 문서에서 자주 등장하는 단어는 중요도가 낮다고 판단
ex) The
- 특정 문서에서만 자주 등장하는 단어는 중요도가 높다고 판단

TFIDF의 약점

- If core term/concept isn't actually used much in document
- Poor searches

02 TFIDF

Variants on TF

- 0/1 Boolean frequencies
- Logarithmic frequencies $\rightarrow \log(\text{TF}+1)$
- Normalized frequency

02 TFIDF

TFIDF 실제로는 더욱 복잡하다

- N-grams ex) Computer Science \neq Computer + Science
- Significance in Documents
- General Document Authority
- Implied Content



01 Introduction



02 TFIDF



03 Content Based
Filtering

03 Content Based Filtering

Keyword Vector

- Each keyword is a dimension
- Each item has a position → vector
- Each user has a profile → vector
- Match : how closely the two vectors align?
- Limit keyword space ex) stem and stop

03 Content Based Filtering

Formalization

$$t \in T$$

T_i : the set of tags applied to item i

t_{ui} : *tag application*

\rightarrow_{t_i} : *Weighted vector of tags*

03 Content Based Filtering

Prediction

- Profile vector와 Item vector간의 Cos 계산
- 범위 : $[-1, 1]$
- 같은 방향 \rightarrow 최댓값 1
- Top-n : 상위 n개의 데이터를 사용하여 Predict

03 Content Based Filtering

이러한 접근 방식의 이점

- Content-Based
- Understandable Profile
 - 수정이 가능
- Easy Calculation

이러한 접근 방식의 한계점

- 적절한 Weight와 Factor
 - Reiteration에 주의
- Interdependencies



Thank you :)

들어 주셔서 감사합니다