ISSN: 0975-766X
CODEN: IJPTFI
Research Article

Available Online through
www.ijptonline.com

ONLINE CELEBRITY RECOGNITION

Mizpah Jenny. D[1], Arul.k[2]
UG scholar[1], Assistant Professor[2],
Department of Computer Science and Engineering,
Saveetha School of Engineering, Saveetha University, Chennai.

Received on: 25.09.2016 *Accepted on: 15.10.2016*

Abstract

In this paper, a online celebrity recognition scheme is presented. The celebrity centre, that includes personal metadata and confidential tags, is crafted from Wikipedia. Celebrity credit ability is endowed to recognize celebrities in articles established on the celebrity base. Two simple demos are gave to display the possible custom of celebrity recognition for personalized recommendation and smart browsing.

Keyword: Celebrity recognition, Meta data, Personalized recommendation, Smart browsing.

1. Introduction

The crucial task of personalized recommendation is to predict what a user is really concerned by collecting and scanning user's history data. For example, a user browsed Web pages often mention brazil players like, Rolando and rolandhino. By scanning these appeared names with the data mining techniques, it is easy to know the user is a football fan, his favourite team is brazil, and his favourite player might be Rolando or rolandhino. Clearly, such kinds of information have immense help for personalized recommendation. When browsing Web pages, it's often to notice some unknown or unfamiliar concepts. It will be very supportive and suitable if the browser can automatically detect and highlight them, and further shows an explanation window when mouse hovered over them. We call such a technique smart browsing, which helps a users surfing the Web like an intelligent supporter. Clearly, recognizing celebrity names is important for smart browsing. However, a celebrity base which contains the knowledge of a vast amount of celebrities is the basic of celebrity recognition. Fortunately, Wikipedia is a ready-made source for constructing celebrity base. Wikipedia has more than 2.2million English articles as of March 2008, and a reasonable portion of them are biographies documenting persons especially the celebrities in history or at the present time. As shown in Figure 1, a Wikipedia-based online celebrity recognition scheme is proposed. The celebrity foundation is constructed from Wikipedia during the construction phase. Celebrity recognition service is given to recognize celebrities in articles based on the celebrity foundation. Applications are enhanced by the celebrity recognition service. The features of Figure 1 are described in the following sections.

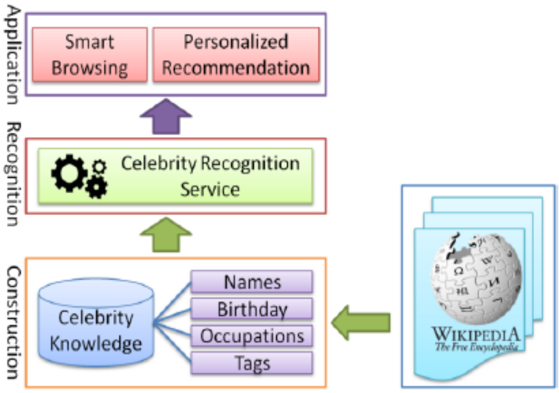


Figure 1. Scheme of Wikipedia-based Celebrity Recognition

2. Celebrity Based Construction

The counselled celebrity centre includes confidential metadata and confidential tags. Confidential metadata are confidential information such as terms, birthday and occupations that can be extracted from Wikipedia articles directly. Confidential tags are words or phrases that associated alongside a people. Usually, the tags are disparate from person to person due to the exceptional person experiences. For examples, tags for Abraham Lincoln could be “president of the United States”, “American war” and “bondage”, but the tags for Bill Gates might be “world's wealthiest people”, “operating system” and “Microsoft”. During the celebrity core assembly period, we have used Gate [Cunningham] to comprehend a little natural language processing (NLP) purposes, like part-of-speech (POS) tagging, noun phrase trunking (NPT) and outline matching. The features are debated in the following.

3. Profile Article Finding

At the extremely early, we demand to find out profile articles from all Wikipedia articles. Totally 295,693 profile articles are discovered by the pursuing two easy heuristic rules:

1) Profile articles are always categorized as “living people” or “person data”.

2) Profile articles are always started alongside a description in a little distinct outlines, such as “"'XXX YYY'" (born [[June 12]], [[1924]]) …”.

3. Confidential Metadata Extraction The target of confidential metadata extraction is to remove person’s metadata data, such as terms, birthday and occupations from profile articles. The profile articles have two advantaged characteristics for confidential metadata extraction:

1) Usually, the early paragraph exceptionally the early sentence of a profile article is the summarization of the person, that frequently encompasses data such as terms, birthday and occupations;

2) Moreover, a little profile articles contain an info box, that catalogues the confidential metadata in structured table, so metadata in info box are facile to extract.

4. Name Extraction

The label of a profile article is normally the term of the person. Though, terms removed from label, info box and the article could be different, such as “john F kennedy”, “JFK” and “John Fitzgerald Kennedy”. A “Full Name” is selected from all removed terms by heuristic rules. The other removed terms are marked as “Alias Name”. Moreover, the “Full Name” is separated to “First Name”, “Middle-Name”, “Last Name” and “Suffix Name”.

5. Birthday Extraction

In a profile article, the birthday is frequently in paired parentheses following the terms, or right afterward words such as “born” or “be born”. But birthdays could be composed in so many forms, such as “June 12, 1924”, “12 June, 1924”, “Jun. 12, 1924”, “June, 1924”, “1924” or “BC 124”. Pattern rules are crafted to parse birthdays in disparate forms.

6. Occupation Extraction

Compare alongside term extraction and birthday extraction, occupation extraction is far extra tough, because occupations are expressed in nature speech of free styles.

Prior to the extraction, an occupation catalogue (about 1500 occupations) is generated from Wikipedia: http://en.wikipedia.org/wiki/List_of_jobs. It’s obviously that not all the words matched in profession catalogue in the early paragraph are professions of the person, for example the word “general” could be an adjective means “common”.

To ascertain this, POS and NPT data are introduced. The detail algorithm is as follows:

1) Mark the income words and NPTs in the before paragraph;

2) Filter occupation words of that POS are not noun;

3) Filter occupation words not in a NPT; 4) Match kept occupation words alongside predefined description outlines

7. Confidential Tag Summarization

The target of confidential tags summarization is to summarize personal tags from profile articles.

8. Tag Generation

A confidential tag is an n-gram of words(namely, a thread of words) that appears in the article. Allow $t = w1w2 \dots wN$ represent a tag t that encompassed N words.

Clearly, not all n-grams are reasonable confidential tags. As Table 1 shows, the 1-gram “the” (which is a usually utilized word and no specific meaning) and the 2-gram “was elected” (no more information than “elected”) are not reasonable tags.

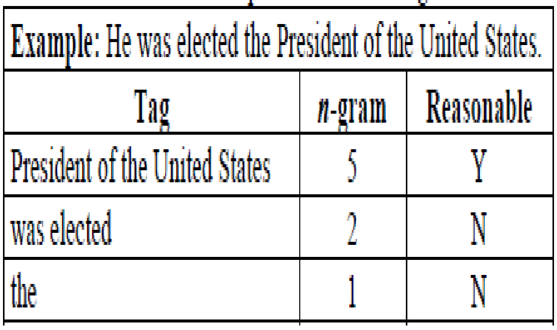


Table 1. Examples of Personal Tags

Example: He was elected the President of the United States.		
Tag	*n*-gram	Reasonable
President of the United States	5	Y
was elected	2	N
the	1	N

Therefore, it’s vital to filter unreasonable tags.

Term frequency (TF), document frequency (DF) and POS information are used. Here, TF t, , TF t and DF t represent the emergence number of t in article d, the appearance number of t in all Wikipedia articles, and the number of Wikipedia articles that encompass t, respectively. Firstly, n-grams alongside low TF worth or low DF worth are filtered. In our examinations, filter conditions are set to TF t, d< 2 or TF t ≤ 2 or DF t< 2.Secondly, n-grams matching POS filtering laws are filtered. Currently we have set up 47 rules. A POS filtering rule R is described as:

$$R \rightarrow (n \odot N_0) \wedge \left(\bigwedge_{i=1}^{n} (p_i = P_i) \right)$$

where$N0$ is a affirmative integer, ⊙= <,=,>, ≤, ≥ , n is the number of words to be matched, pi is the POS of wi , and Pi is a POS set. The meaning of POS tags complies with the meaning in Penn Treebank Project. For example, $R39$ → n ≥ 2 ∧pn = JJ is a law implying that if the length of n-gram is equal or larger than 2, and the POS of the last word is “adjective”, next the n gram should be filtered.

9. Tags Ranking

According to human’s intuition, tags that representative and discriminative ought to have elevated ranked value: □ Representative: Tags that can embody the people. Φ t embodies representative degree of t. □ Discriminative: Tags that can differentiate the people and additional peoples. Ψ t embodies discriminative degree of t. Then, the ranking worth of t is described by R

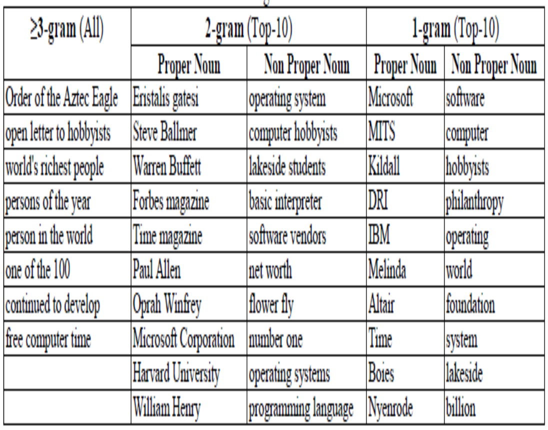
t =Φ t × Ψ . In our examinations, we allow

Φ t = TF t and Ψ t = ln 1 + DF t−1

10. Experimental Result

It's extremely tough to assess the tag summarization performance just by a worth, because whether a tag is better than one more is mainly depends on human's opinion. Therefore, we merely catalogue the ranked confidential tag summarization result. We seize the Wikipedia article for Bill Gates as the example (http://en.wikipedia.org/wiki/Bill_gates). Table 2 displays the removed top-10 1-grams, 2-grams and 3(or above)-grams tags. It's certainly that these tags are meaningful and closed connected to Bill Gates.

Table 2. Personal Tags Summarization Result



≥3-gram (All)	2-gram (Top-10)		1-gram (Top-10)	
	Proper Noun	Non Proper Noun	Proper Noun	Non Proper Noun
Order of the Aztec Eagle	Eristalis gatesi	operating system	Microsoft	software
open letter to hobbyists	Steve Ballmer	computer hobbyists	MITS	computer
world's richest people	Warren Buffett	lakeside students	Kildall	hobbyists
persons of the year	Forbes magazine	basic interpreter	DRI	philanthropy
person in the world	Time magazine	software vendors	IBM	operating
one of the 100	Paul Allen	net worth	Melinda	world
continued to develop	Oprah Winfrey	flower fly	Altair	foundation
free computer time	Microsoft Corporation	number one	Time	system
	Harvard University	operating systems	Boies	lakeside
	William Henry	programming language	Nyenrode	billion

11. Celebrity Reorganization

With the celebrity centre, it is possible to recognize celebrities from a given article. At first, each capitalized word in the given article is believed to be a first name or last name of a celebrity. Then, for each capitalized word, taking all the celebrities, whose first name or last name is the capitalized word, in the celebrity base as candidates. In the end, these candidates are scored according to the knowledge matching result within the contexts around the capitalized word. The celebrity with the greatest score is accepted as the one that the capitalized word refers to.LetScw is the tally of a capitalized word w refers to celebrity c. The score is calculated as

$$S_c(w) = \sum_{T}^{'} \left(\omega_T \times A_T(c, x_T, y_T)\right)$$

$$T = \begin{Bmatrix} \{\text{LastName, FirstName, MiddleName,} \\ \text{Birthday, Occupation, Tag}\} \end{Bmatrix}$$

where ωT is the weight, ATc, lT is the appearance number of elements belongs to T of celebrity c in the contexts (xT words before w and yT words after w). For example, ωTag = 2, xT = 50, and yT = 50.

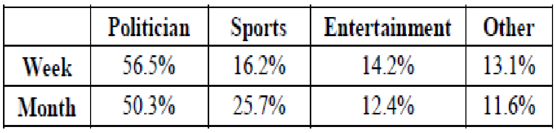
12. Applications Based on Celebrity Recognition

We have consolidated online celebrity credit for Web page browsing. A user script of Grease monkey for Firefox is industrialized to arrest the Web pages user browsing and send the html basis to the celebrity credit service. The ability analyzes the page and knows all the celebrities. The credit consequence can be utilized for personalized recommendation and intelligent browsing.

13. Personalized Recommendation

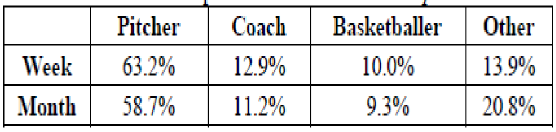
We amass all the celebrities that the user browsed ever, and next present statistic scrutiny and data mining. Table 3 is the occupation class allocation consequence of a user browsed celebrities inside one month. We can find that the user is interested in government most and sport is his second lover. Furthermore, we can scrutiny the occupations among "games". As Table 4 shows, we can find the user's favourite sport games is baseball. Such information will be very useful for further recommendations for the user.

Table 3. Profession Class Distribution



	Politician	Sports	Entertainment	Other
Week	56.5%	16.2%	14.2%	13.1%
Month	50.3%	25.7%	12.4%	11.6%

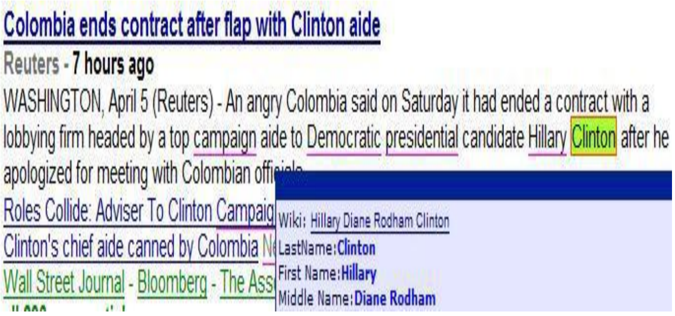
Table 4. Occupation Distribution in *"Sports"*



	Pitcher	Coach	Basketballer	Other
Week	63.2%	12.9%	10.0%	13.9%
Month	58.7%	11.2%	9.3%	20.8%

14. Intelligent Browser

Recognized celebrity can be utilized for intelligent browsing. Figure 2 displays the highlights and popup menus of Web pages in Firefox. The words in rectangles alongside green background are last terms of celebrities, and the words underlined byred lines are the words that prop the celebrities.



Colombia ends contract after flap with Clinton aide

Reuters - 7 hours ago

WASHINGTON, April 5 (Reuters) - An angry Colombia said on Saturday it had ended a contract with a lobbying firm headed by a top campaign aide to Democratic presidential candidate Hillary Clinton after he apologized for meeting with Colombian offi

Roles Collide: Adviser To Clinton Campaig

Clinton's chief aide canned by Colombia N

Wall Street Journal - Bloomberg - The Ass

Wiki: Hillary Diane Rodham Clinton

LastName:Clinton

First Name:Hillary

Middle Name:Diane Rodham

Figure 2. Demo for Smart Browsing

15. Conclusion

In this paper, a Wikipedia-based celebrity recognition scheme is presented. Celebrity knowledge, which includes personal metadata and personal tags, are acquired from Wikipedia. Based on the celebrity base, celebrities in browsing Web pages are recognized. Browsed celebrities history data are collected and analyzed for personalized recommendation. Recognized celebrity names are highlighted and explained for smart browsing. In the future, we plan to fuse person information from different sources to increase current celebrity base.

16. References

1. Magdalini E., Michalis V. 2003. Web Mining for Web Personalization.*ACM Transactions on Internet Technology*, 3(1):1-27. Cunningham H., Maynard D., Bontcheva K., Tablan V. 2002.
2. GATE: A Structural and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings ACL'02*.
3. S. Z. Li and A. K. Jain (eds.), Handbook of Face Recognition, Second edition, Springer, 2011
4. J. Vlahos, "Surveillance Society: New Hig h-Tech Cameras are Watching You," Popular Mechanics, October, 2009
5. L. Ding, C. Shu, C. Fang, and X. Ding, " Computers do better than experts matching faces in a large population," IEEE International Conference on Cognitive Informatics, pp. 280-284, 2010.
6. H. Ling, S. Soatto, N. Ramanathan, and D. Jacobs, "Face verification across age progression using discriminative methods, IEEE Trans. on Information Forensic and Security, Vol. 5, No. 1, pp. 82-91, Mar. 2010.
7. X. Wang and X. Tang, "Face Photo-Sketch Synthesis and Recognition," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 31, No.11, pp.1955-1967, 2009
8. Grease monkey .https://addons.mozilla.org/firefox/748/
9. Dict.cn. http://dict.cn/
10. Clear Forest Gnosis.https://addons.mozilla.org/en-US/firefox/addon/3999