Wikipedia-based Online Celebrity Recognition

Demiao Lin, Jianming Jin, Yuhong Xiong

HP Laboratories

HPL-2010-2

Keyword(s):
celebrity recognition, wikipedia, keyword extraction

Abstract:

In this paper, a Wikipedia-based online celebrity recognition scheme is presented. The celebrity base, which includes personal metadata and personal tags, is constructed from Wikipedia. Celebrity recognition service is provided to recognize celebrities in articles based on the celebrity base. Two simple demos are introduced to show the potential usage of celebrity recognition for personalized recommendation and smart browsing.

External Posting Date: January 6, 2010 [Fulltext] Approved for External Publication
Internal Posting Date: January 6, 2010 [Fulltext]

Wikipedia-based Online Celebrity Recognition

Demiao Lin, Jianming Jin, Yuhong Xiong

Hewlett-Packard Labs, China
Tower A505 SP Tower, Tsinghua Science Park, HaiDian District, Beijing, China, 100084
{demiao.lin, jian-ming.jin, yuhong.xiong}@hp.com

Abstract

In this paper, a Wikipedia-based online celebrity recognition scheme is presented. The celebrity base, which includes personal metadata and personal tags, is constructed from Wikipedia. Celebrity recognition service is provided to recognize celebrities in articles based on the celebrity base. Two simple demos are introduced to show the potential usage of celebrity recognition for personalized recommendation and smart browsing.

Introduction

The crucial task of personalized recommendation [Magdalini] is to predict what a user is really concerned by collecting and analyzing user's history data. For example, a user browsed Web pages often mention NBA players, Tracy McGrady and Ming Yao. By analyzing these appeared names with some data mining techniques, it is easy to know the user is a NBA fan, his favorite team is Houston Rockets, and his favorite player might be Tracy McGrady or Ming Yao. Clearly, such kinds of information have great help for personalized recommendation.

When browsing Web pages, it's often to notice some unknown or unfamiliar concepts. It will be very helpful and convenient if the browser can automatically detect and highlight them, and further pop up an explanation window when mouse hovering over them. We call such a technique smart browsing, which helps users surfing the Web like an intelligent assistant. Obviously, recognizing celebrity names is important for smart browsing.

However, a celebrity base which contains the knowledge of a huge amount of celebrities is the basic of celebrity recognition. Fortunately, Wikipedia is a ready-made source for building celebrity base. Wikipedia has more than 2.2 million English articles as of March 2008, and a quite portion of them are biographies documenting persons especially the celebrities in history or nowadays.

As shown in Figure 1, a Wikipedia-based online celebrity recognition scheme is proposed. The celebrity base is constructed from Wikipedia during the construction phase. Celebrity recognition service is provided to recognize celebrities in articles based on the celebrity base. Applications are enhanced by the celebrity recognition service. The details of Figure 1 are discribed in the following sections.

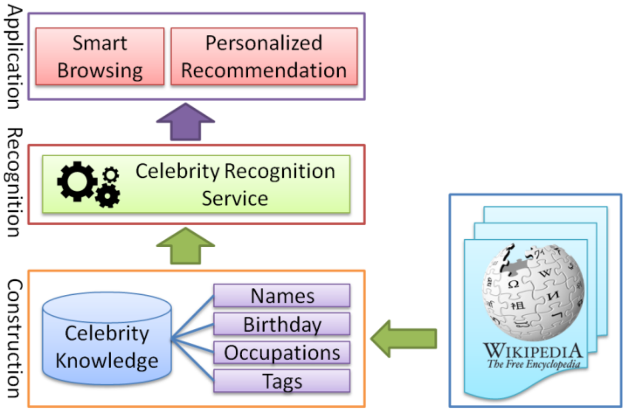


Figure 1. Scheme of Wikipedia-based Celebrity Recognition

Celebrity Base Construction

The proposed celebrity base includes personal metadata and personal tags. Personal metadata are personal information such as names, birthday and occupations which can be extracted from Wikipedia articles directly. Personal tags are words or phrases which associated with a people. Usually, the tags are different from person to person due to the unique person experiences. For examples, tags for Abraham Lincoln might be "*president of the United States*", "*American civil war*" and "*slavery*", but tags for Bill Gates might be "*world's richest people*", "*operating system*" and "*Microsoft*".

During the celebrity base construction phase, we have used Gate [Cunningham] to realize some nature language processing (NLP) functions, such as part-of-speech (POS) tagging, noun phrase trunking (NPT) and pattern matching. The details are discussed in the following.

Biography Articles Finding

At the very first, we need to find out biography articles from all Wikipedia articles. Totally 295,693 biography articles are found by the following two simple heuristic rules: 1) Biography articles are always categorized as "*living people*" or "*person data*". 2) Biography articles are always started with a description in some special patterns, such as "*'''XXX YYY''' (born [[June 12]], [[1924]]) …*".

Personal Metadata Extraction

The aim of personal metadata extraction is to extract person's metadata information, such as names, birthday and occupations from biography articles.

The biography articles have two advantaged characteristics for personal metadata extraction: 1) Usually, the first paragraph especially the first sentence of a biography article is the summarization of the person, which often contains information such as names, birthday and occupations; 2) Moreover, some biography articles include an infobox, which lists the personal metadata in structured table, so metadata in infobox are easy to extract.

Name Extraction

The title of a biography article is usually the name of the person. However, names extracted from title, infobox and the article may be different, such as "*George W. Bush*", "*George Bush*" and "*George Walker Bush*". A "*FullName*" is selected from all extracted names by heuristic rules. The other extracted names are marked as "*AliasName*". Moreover, the "*FullName*" is split to "*FirstName*", "*MiddleName*", "*LastName*" and "*SuffixName*".

Birthday Extraction

In a biography article, the birthday is often in paired parentheses following the names, or right after words such as "*born*" or "*be born*". But birthdays may be written in so many forms, such as "*June 12, 1924*", "*12 June, 1924*", "*Jun. 12, 1924*", "*June, 1924*", "*1924*" or "*BC 124*". Pattern rules are created to parse birthdays in different forms.

Occupation Extraction

Compare with name extraction and birthday extraction, occupation extraction is much more difficult, because occupations are expressed in nature language of free styles.

Prior to the extraction, an occupation list (about 1500 occupations) is generated from Wikipedia article: http://en.wikipedia.org/wiki/List_of_occupations. It's obviously that not all the words matched in occupation list in the first paragraph are occupations of the person, for example the word "*general*" may be an adjective means "*common*". To resolve this, POS and NPT information are introduced. The detail algorithm is as follows: 1) Mark the occupation words and NPTs in the first paragraph; 2) Filter occupation words of which POS are not noun; 3) Filter occupation words not in a NPT; 4) Match reserved occupation words with predefined description patterns.

Personal Tags Summarization

The aim of personal tags summarization is to summarize personal tags from biography articles.

Tags Generation

A personal tag is an *n*-gram of words (namely, a string of words) which appears in the article. Let $t = w_1 w_2 \dots w_N$ represent a tag *t* which comprised *N* words. Clearly, not all *n*-grams are reasonable personal tags. As Table 1 shows, the 1-gram "the" (which is a commonly used 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

<table>
<tr><td colspan="3"><b>Example:</b> He was elected the President of the United States.</td></tr>
<tr><th>Tag</th><th><i>n</i>-gram</th><th>Reasonable</th></tr>
<tr><td>President of the United States</td><td>5</td><td>Y</td></tr>
<tr><td>was elected</td><td>2</td><td>N</td></tr>
<tr><td>the</td><td>1</td><td>N</td></tr>
</table>

Therefore, it's necessary to filter unreasonable tags. Term frequency (TF), document frequency (DF) and POS information are used. Here, $\mathrm{TF}(t,d)$, $\mathrm{TF}(t)$ and $\mathrm{DF}(t)$ represent the appearance number of *t* in article *d*, the appearance number of *t* in all Wikipedia articles, and the number of Wikipedia articles which contain *t*, respectively.

Firstly, *n*-grams with low TF value or low DF value are filtered. In our experiments, filter conditions are set to $\mathrm{TF}(t,d) < 2$ or $\mathrm{TF}(t) \leq 2$ or $\mathrm{DF}(t) < 2$.

Secondly, *n*-grams matching POS filtering rules are filtered. Currently we have set up 47 rules. A POS filtering rule *R* is defined as:

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

, where N_0 is a positive integer, $\circledcirc = \{<, =, >, \leq, \geq\}$, n is the number of words to be matched, p_i is the POS of w_i, and P_i is a POS set. The meaning of POS tags complies with the definition in Penn Treebank Project.

For example, $R_{39} \rightarrow (n \geq 2) \wedge (p_n = \{\mathrm{JJ}\})$ is a rule implying that if the length of *n*-gram is equal or greater than 2, and the POS of the last word is "*adjective*", then the *n*-gram should be filtered.

Tags Ranking

According to human's intuition, tags which representative and discriminative should have high ranked value:

- Representative: Tags which can represent the people. $\Phi(t)$ represents representative degree of *t*.
- Discriminative: Tags which can discriminate the people and other peoples. $\Psi(t)$ represents discriminative degree of *t*.

Then, the ranking value of *t* is defined by $\mathrm{R}(t) = \Phi(t) \times \Psi(t)$. In our experiments, we let $\Phi(t) = \mathrm{TF}(t)$ and $\Psi(t) = \ln\left(1 + \mathrm{DF}(t)\right)^{-1}$.

Experiment Result

It's very difficult to evaluate the tag summarization performance just by a value, because whether a tag is better than another is largely depends on human's opinion. Therefore, we only list the ranked personal tag summarization result. We take the Wikipedia article for Bill Gates as the example (http://en.wikipedia.org/wiki/Bill_gates). Table 2 shows the extracted top-10 1-grams, 2-grams and 3(or above)-grams tags. It's obviously that these tags are meaningful and closed related 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

Celebrity Recognition

With the celebrity base, it is feasible to recognize celebrities from a given article. At first, each capitalized word in the given article is supposed to be a first name or last name of a celebrity. Then, for each capitalized word, taking all celebrities, whose last name or first name is the capitalized word, in the celebrity base as candidates. Finally, these candidates are scored according to the knowledge matching result within the contexts around the capitalized word. The celebrity with the maximum score is recognized as the one that the capitalized word refers to.

Let $S_c(w)$ is the score of a capitalized word w refers to celebrity c. The score is calculated as

$$S_c(w) = \sum_T \left(\omega_T \times A_T(c, x_T, y_T)\right)$$

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

where ω_T is the weight, $A_T(c, l_T)$ is the appearance number of elements belongs to T of celebrity c in the contexts (x_T words before w and y_T words after w). For example, $\omega_{Tag} = 2$, $x_T = 50$, and $y_T = 50$.

Applications Based on Celebrity Recognition

We have integrated online celebrity recognition for Web page browsing. A user script of Greasemonkey for Firefox is developed to capture the Web pages user browsing and send the html source to the celebrity recognition service. The service analyzes the page and recognizes all the celebrities. The recognition result can be used for personalized recommendation and smart browsing.

Personalized Recommendation

We collect all the celebrities that the user browsed ever, and then perform statistic analysis and data mining. Table 3 is the profession class distribution result of a user browsed celebrities within one month. We can find that the user is interested in politics most and sport is his second lover. Furthermore, we can analysis the occupations among "Sports". As Table 4 shows, we can find the user's favorite sport is baseball. Such information will be very useful for further recommendations for the user.

Table 3. Profession Class Distribution

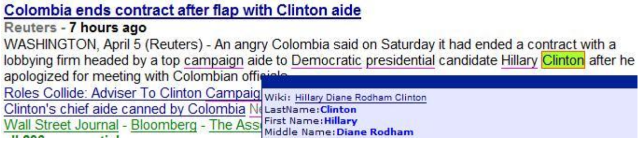
	Politician	**Sports**	**Entertainment**	**Other**
Week	56.5%	16.2%	14.2%	13.1%
Month	50.3%	25.7%	12.4%	11.6%

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%

Smart Browsing

Recognized celebrity can be used for smart browsing. Figure 2 shows the highlights and popup menus of Web pages in Firefox. The words in rectangles with green background are last names of celebrities, and the words underlined by red lines are the words that support 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

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.

References

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. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings ACL'02*.