Estimating the First Appearance of an Entity Mentioned on the Web

Yue Guo

March 15, 2011

Contents

List of Figures

List of Tables

Abstract

News entities such as newly released products or movies are very likely to be mentioned or reported on the Web nowadays. Once a piece of such news appeared on the Web, it could be copied, modified, and republished by many others. Finding the first appearance of the news is desirable under specific circumstances. In this article, an approach based on Google's timeline operation is proposed to the identify the earliest dates of given news entities along with an evaluation of the prototype.

Chapter 1

Introduction

Currently the Web has become one of the largest information sources with all kinds of information of high timeliness on portals, forums, and blogs. Information can be reproduced very quickly by others once it is published. As the information spreads widely, the original source of the information is of great interest when it comes to evaluating the credibility of the information. In addition , other sophisticated tools aiming at discovering the news about the most recent products, such as WebKnox[DU08], need a way to evaluate their timeliness of how quickly they are able to capture information of new products. This could be measured by calculating the interval between the date when they capture the information of a product and the date when a piece of information of that product appears for the first time on the Web. In this scenario, finding the source of information is critical.

However, not every reproduction of a piece of information has the backlink to the original publisher, which makes the source difficult to locate. On the other hand, it is not easy for computer programs to tell whether a link is the exact backlink of the information source without clear semantics that computers understand. Moreover, it is difficult for computers to tell when a piece of information is released as humans do, also because of the lack of the necessary semantics for the web page contents. Therefore, how to identify the source of certain information is worth researching very much.

In this article, an approach based on Google's timeline operation to achieve this goal is presented. A prototype called "FirstApp Finder" is implemented and evaluated as well. The structure of this article is organized as follows: in Section 2, background knowledge is introduced. Section 3 presents the details of how to estimate the date of the first appearance on the Web of a certain entity. Results of the evaluation is shown in Section 4. Section 5 gives an outlook of future work and Section 6 draws a conclusion.

Chapter 2

Background

One prime step to find the earliest appearance of a news entity is to find the date of all its appearances somewhere. Because computers are not as intelligent as humans, they can not understand the contents of web pages, not to mention to find the publish date of certain information. There has to be some tools which can be relied on to fulfill the task. In this section, some basic knowledge as well as the toolkit used in the implementation will be introduced.

2.1 Where to Search

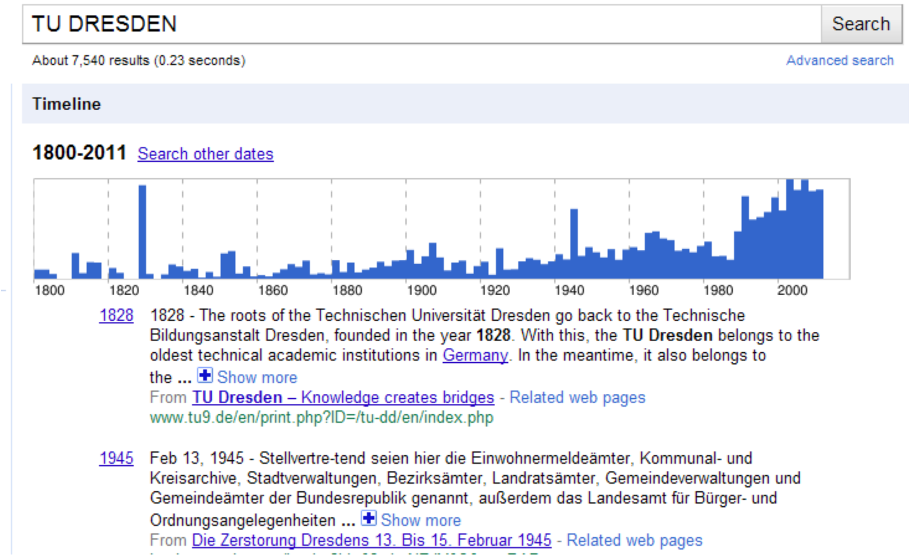
There are many portals and forums dedicated to certain fields, such as IT products, movies, cars, etc. They are major information sources and one can expect they have the most up-to-date first hand news regarding specific area. Looking for the earliest appearance of an entity in such websites is easier due to the limited scope and the search functionalities they provide. However, the whole Web instead of fixed information sources is selected as where to search the first appearance of a certain entity. If fixed information source is used, whenever searching the first appearance of an entity which does not belong to any category of these sources, one has to introduce a new kind of information source, which is neither flexible nor convenient.

2.2 Google's Timeline Operation

Google is chosen as the research object as it is currently one of the best search engine in the world. The most important reason is Google provides a new tool called "Timeline", which is the foundation of the approach presented in this article. What makes it different from a regular Google search operation? As shown in Figure 2.1, Google's timeline operation concludes the "history" of a keyword by revealing all the dates related to the keyword being searched. The height of the bars in the chart reflects how close the relevance between the keywords being searched and that year is. The more times a year is mentioned with the keywords, the higher the bar is. When clicking on the bars, a new chart for every month in the corresponding year is shown. The most relevant pages containing the keywords are also listed. This is critical to our work, since it is very difficult to find all appearance dates through a regular Google search because of too many redundant dates in the result set. How Google's timeline operation can be utilized to find the first appearance of an entity will be discussed in Section 3.

2.3 TUD Palladian Toolkit

Palladian is a collection of algorithms for text processing focused on classification, extraction, and retrieval[DU11]. It is used to evaluate the age of a page and also as a language filter. In this article,



About 7,540 results (0.23 seconds)

1800-2011

1828 1828 - The roots of the Technischen Universität Dresden go back to the Technische Bildungsanstalt Dresden, founded in the year **1828**. With this, the **TU Dresden** belongs to the oldest technical academic institutions in Germany. In the meantime, it also belongs to the ... ➕ Show more
From **TU Dresden** – Knowledge creates bridges - Related web pages
www.tu9.de/en/print.php?ID=/tu-dd/en/index.php

1945 Feb 13, 1945 - Stellvertre-tend seien hier die Einwohnermeldeämter, Kommunal- und Kreisarchive, Stadtverwaltungen, Bezirksämter, Landratsämter, Gemeindeverwaltungen und Gemeindeämter der Bundesrepublik genannt, außerdem das Landesamt für Bürger- und Ordnungsangelegenheiten ... ➕ Show more
From Die Zerstorung Dresdens 13. Bis 15. Februar 1945 - Related web pages

Figure 2.1: TU Dresden's Timeline

the first appearance of an entity means its first appearance on English or German webpages for simplicity. However, the filtering can be turned off if necessary.

Chapter 3

Related Work

There are several prior researches regarding estimating source dates of entities. [BC09] presents a source date estimation algorithm using the longest dense sequence in a set of dates for a news entity. This approach, however, relies much on the constant used as a threshold to find the longest dense sequence, which is difficult to uniformly define for different entities. Besides, as to entities such as products or movies, rumors or previews often appear on the Web before their final release. The longest dense usually is too late compared to our goal, i.e. finding the earliest mention of a news entity. In [BC09], the experiment of finding source dates has a time range from October, 2003 to July, 2006 for the queries. This range is extended in our work from 1990 to present, since our goal is to find the earliest mention of an entity as much as possible. On the other hand, [BC09] only used **Last-Modified** keyword in HTTP header to evaluate the estimated dates in the experiment. In our work, we used TUD Palladian toolkit to evaluate page age to the most extent via different technologies. Other researches, e.g. temporal summarization[JAK01] and timeline construction[SJ00] has more similar intent to Google's timeline, instead of finding the earliest mention of an entity through a timeline, which is the method we use to achieve our goal.

Chapter 4

Estimate the First Appearance of an Entity

This section mainly presents the details of how FirstApp Finder is implemented, including the algorithm for estimating the appearance year of an entity, the assumptions the algorithm is based on and how the final result is generated.

4.1 Assumptions

Through observation, the height of the bars in the yearly timeline chart is reflects when an entity goes in public in most cases. Because the height of bars usually represents the relevance between the year and the keyword. The highest bar mostly means that is when an entity is officially released or published. This is especially true for products, movies and whatever is used to be new to people. Examples are depicted in Figure 4.1 and 4.2. One is easy to notice that the years with the highest bars, 2010 and 2006, are exactly when the tablet computer iPad and the movie Blood Diamond was released. On the other hand, rumors or previews of these entities are very common, therefore, it is obvious that the first appearance of such entities must be sometime in or before the year with the highest bar. As shown in the table in the evaluation section, most of entities, including the two above mentioned entities "iPad" and "blood diamond", can be found first mentioned much earlier than their release time.

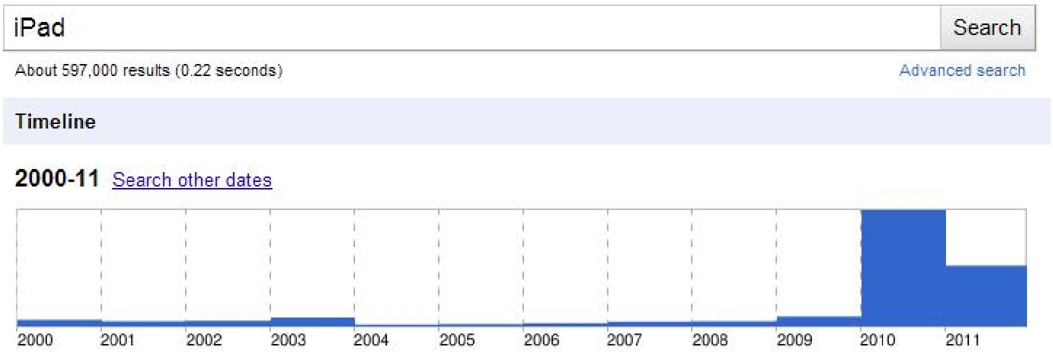
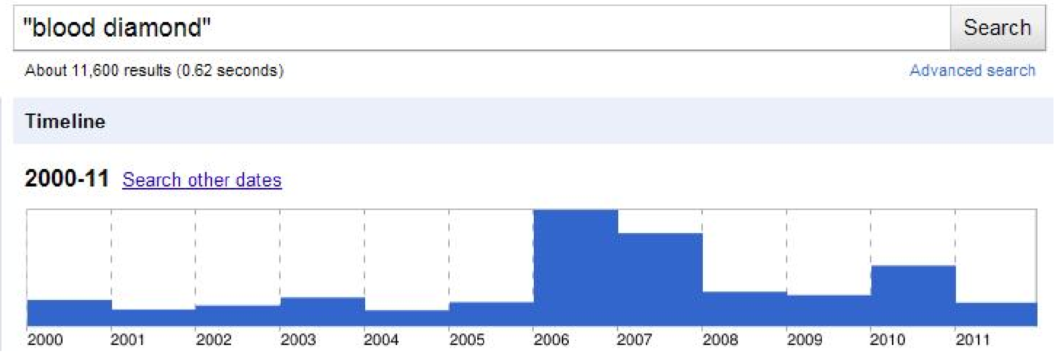


Figure 4.1: Timeline of iPad, Released in 2010

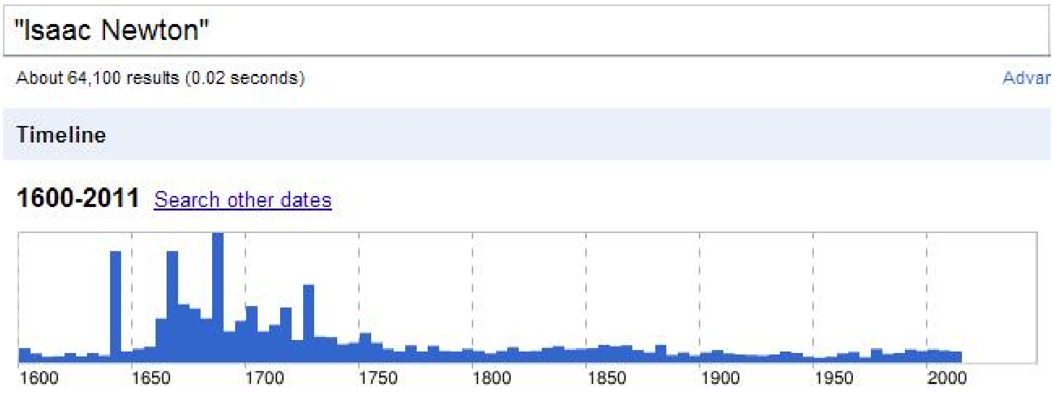
However, this assumption does not hold when it comes to historical events or people, because the highest bars will be in the years when the events happened or the people lived. For example, when "Isaac Newton" is searched, the timeline of it is as shown in Figure 4.3.



"blood diamond"

About 11,600 results (0.62 seconds)

Figure 4.2: Timeline of the 2006 Movie ”Blood Diamond”



"Isaac Newton"

About 64,100 results (0.02 seconds)

Figure 4.3: Timeline of Isaac Newton

When taking only a part of the timeline into consideration, e.g. timeline after 1990(about when WWW went public), the assumption also holds(Figure 4.4).

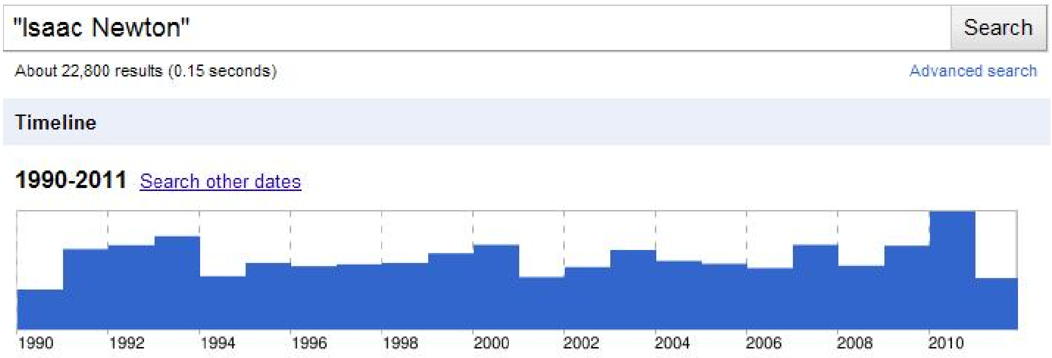


Figure 4.4: Timeline of Isaac Newton from 1990

As this article mainly serves systems such as WebKnox, which aims at capturing most recent products and other news items instead of historical entities decades ago, the assumption that the first appearance of an entity is sometime in or before the year with the peak bar is proven in the evaluation to be safe(see Section 5). The search range of timeline is set to 1990 till the current year, since the dates before the the use of WWW are pointless.

4.2 Problems With Google's Timeline

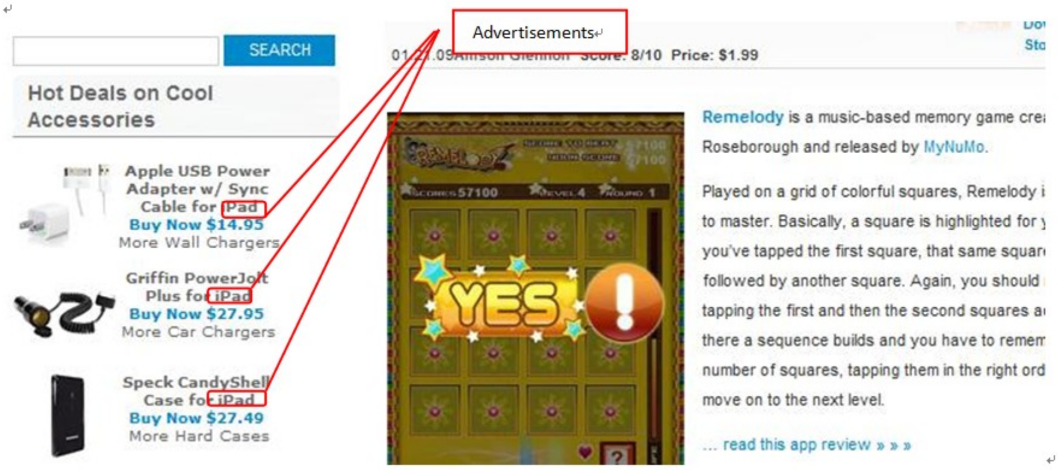
Google's timeline operation is not 100% reliable. The most obvious problem of Google timeline is that not every date shown in the chart and result set is really relevant to the keywords being searched. It is because Google timeline detects all dates which are close to the the keywords being searched in a webpage instead of really understand what these dates mean to the keywords. There are several kinds of dates which impedes FirstApp Finder to make correct estimations to a large extent.

Firstly, dynamic advertisements on webpages are a big obstacle as to the search of the first appearance of an entity. This is especially true for new products. These ads usually lead to a false high relevance of a year which is far earlier from when the product was firstly mentioned. Figure 4.5 shows an example of dynamic loaded ads on a page. The page is inferred as relevant to iPad, however, it is an old page having nothing to do with iPad apart from these ads.

Secondly, adjacent irrelevant dates around the keyword being searched on webpages also interferes the result of timeline. An example is shown in Figure 4.6.

Thirdly, there is a high possibility that Google takes a registration date of a forum user as a date relevant to the keyword, as shown in Figure 4.7. This kind of mistakes lead to results which are too early, since a user's registration date is always earlier than the actual post date of the content with the keyword.

Last but not least, identical entity names with different meanings also result in incorrect relevance between dates and entities. Figure 4.8 is an example where the timeline shows the entity "Spy Next Door" first appeared in 2007, however, the article is in fact not talking about the Jackie Chan movie we want to search. The solutions to these problems will be discussed with other implementation details in the next section.


Advertisements
SEARCH
Hot Deals on Cool Accessories
Dov
Sto
01/21.09 Allison Glennon Score: 8/10 Price: $1.99

Apple USB Power Adapter w/ Sync Cable for iPad
Buy Now $14.95
More Wall Chargers

Griffin PowerJolt Plus for iPad
Buy Now $27.95
More Car Chargers

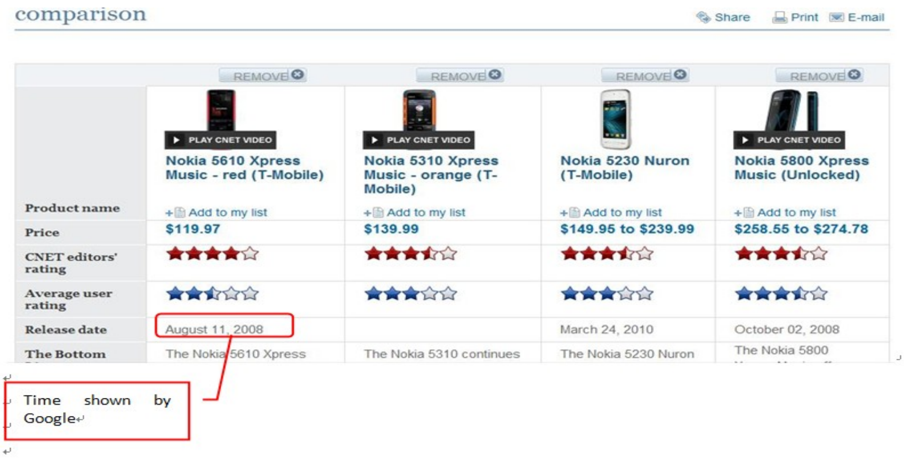
Speck CandyShell Case for iPad
Buy Now $27.49
More Hard Cases

Remelody is a music-based memory game crea
Roseborough and released by MyNuMo.

Played on a grid of colorful squares, Remelody i
to master. Basically, a square is highlighted for y
you've tapped the first square, that same square
followed by another square. Again, you should
tapping the first and then the second squares a
there a sequence builds and you have to remem
number of squares, tapping them in the right ord
move on to the next level.

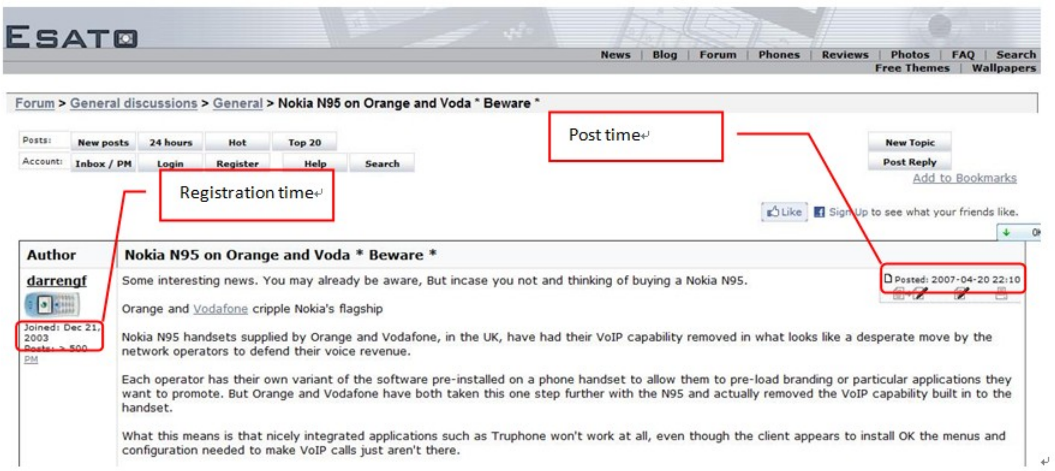
... read this app review » » »

Figure 4.5: Dynamic Ads on Webpages



	REMOVE	REMOVE	REMOVE	REMOVE
Product name	Nokia 5610 Xpress Music - red (T-Mobile) + Add to my list	Nokia 5310 Xpress Music - orange (T-Mobile) + Add to my list	Nokia 5230 Nuron (T-Mobile) + Add to my list	Nokia 5800 Xpress Music (Unlocked) + Add to my list
Price	$119.97	$139.99	$149.95 to $239.99	$258.55 to $274.78
CNET editors' rating	★★★★☆	★★★½☆	★★★½☆	★★★½☆
Average user rating	★★½☆☆	★★★☆☆	★★★☆☆	★★★½☆
Release date	August 11, 2008		March 24, 2010	October 02, 2008
The Bottom	The Nokia 5610 Xpress	The Nokia 5310 continues	The Nokia 5230 Nuron	The Nokia 5800

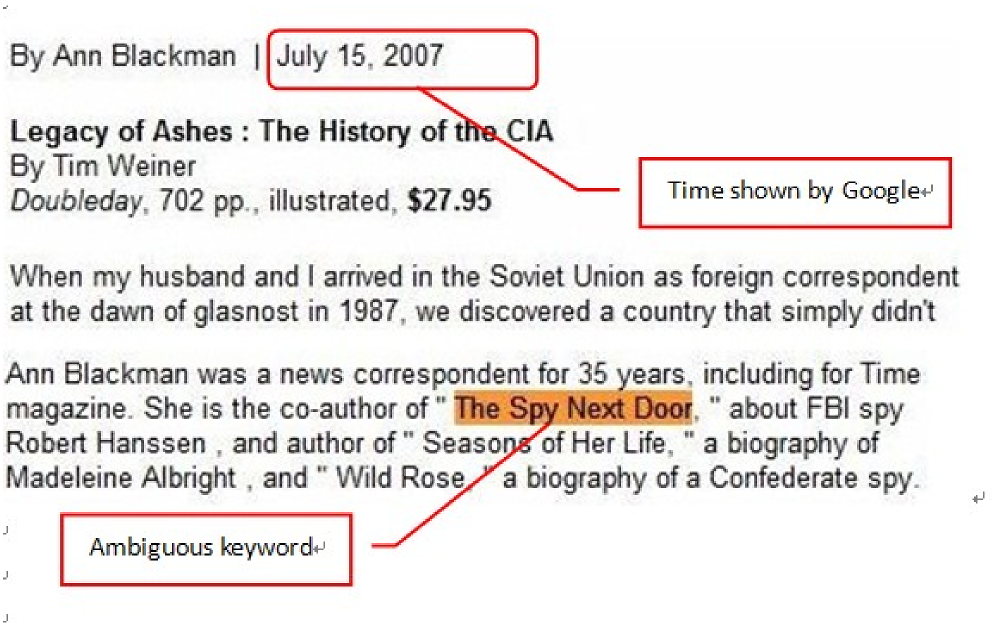
Figure 4.6: Adjacent Irrelevant Dates


ESATO
News | Blog | Forum | Phones | Reviews | Photos | FAQ | Search
Free Themes | Wallpapers
Posts: New posts 24 hours Hot Top 20
Account: Inbox / PM Login Register Help Search
Post time
Registration time
New Topic
Post Reply
Add to Bookmarks
Like
Sign Up to see what your friends like.

Forum > General discussions > General > Nokia N95 on Orange and Voda * Beware *

Author	Nokia N95 on Orange and Voda * Beware *
darrengf Joined: Dec 21, 2003 Posts: > 500 PM	Posted: 2007-04-20 22:10 Some interesting news. You may already be aware, But incase you not and thinking of buying a Nokia N95. Orange and Vodafone cripple Nokia's flagship Nokia N95 handsets supplied by Orange and Vodafone, in the UK, have had their VoIP capability removed in what looks like a desperate move by the network operators to defend their voice revenue. Each operator has their own variant of the software pre-installed on a phone handset to allow them to pre-load branding or particular applications they want to promote. But Orange and Vodafone have both taken this one step further with the N95 and actually removed the VoIP capability built in to the handset. What this means is that nicely integrated applications such as Truphone won't work at all, even though the client appears to install OK the menus and configuration needed to make VoIP calls just aren't there.

Figure 4.7: User Registration Dates Interferes



By Ann Blackman | July 15, 2007

Legacy of Ashes : The History of the CIA
By Tim Weiner
Doubleday, 702 pp., illustrated, **$27.95**

When my husband and I arrived in the Soviet Union as foreign correspondent at the dawn of glasnost in 1987, we discovered a country that simply didn't

Ann Blackman was a news correspondent for 35 years, including for Time magazine. She is the co-author of " The Spy Next Door, " about FBI spy Robert Hanssen , and author of " Seasons of Her Life, " a biography of Madeleine Albright , and " Wild Rose, " a biography of a Confederate spy.

Figure 4.8: "Spy Next Door" Here is Not the Jackie Chan Movie

4.3 Implementation Details

4.3.1 Year Estimation

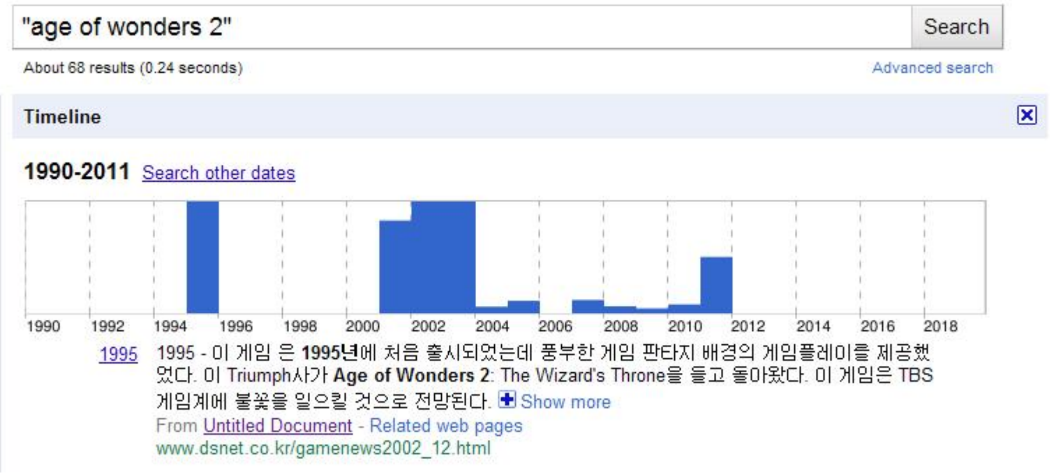
One straightforward solution to the above mentioned problems of timeline is to use a website blacklist to filter those websites which often lead to false results. Using more specific keywords, e.g. an entity name together with the category it belongs to when searching is also a good method to reduce false results.

The estimation of first appearance year is mainly based on the assumption made above, thus check the years before the year with the highest bar. However, because of the flaws of Google's timeline previously discussed, simply finding the highest bar does not always work. Some adjustments are necessary in order to make more precise estimations. Through observations, there are some patterns in the timeline bar chart which tell if a peak bar (a very high bar instead of the highest one) is correct or not. For example, if a peak has no successive bars in the next year slot, it is a false peak, since in most cases, there will not be zero pages regarding an entity after it draws people's attention, even though sometimes the height of the bars next to the peak bar drops dramatically, which suggests there are much less pages mentioning the entity as time passes. An example is shown in Figure 4.9. The PC game "Age of Wonders 2" was actually released in 2002 and was first mentioned in 2001. However, just because the year 1995 is mentioned in a very close position to this keyword on a Korean website, a peak bar appears in 1995.

Another common pattern is that when there are multiple peaks followed by non-zero height bars in the chart, the first one is the most possible date proximate to the first appearance date. This could be explained as the entity somehow is still after its first appearance. An example is shown in Figure 3.10. The famous online game "World of Warcraft" becomes even hotter after its first release in 2004 because of the following version patches of new contents.

Besides, a year in which its bar is with a height less than 10% of the maximum height[1] is in most cases unlikely to be the first appearance year of an entity, unless it is the first year before a high bar (higher than 80%), because there often some rumors or preview news about an entity(especially products or movies) before its release. This is where the search for the first appearance should stop.

[1] How the height of a bar is determined is described here: Google Chart API Documentation.



1995 1995 - 이 게임 은 **1995년**에 처음 출시되었는데 풍부한 게임 판타지 배경의 게임플레이를 제공했었다. 이 Triumph사가 **Age of Wonders 2**: The Wizard's Throne을 들고 돌아왔다. 이 게임은 TBS 게임계에 불꽃을 일으킬 것으로 전망된다. ✚ Show more

From Untitled Document - Related web pages

www.dsnet.co.kr/gamenews2002_12.html

Figure 4.9: False Peak

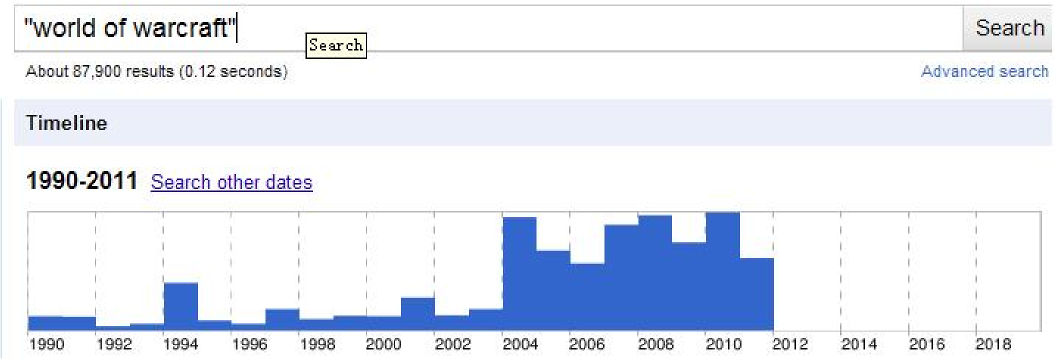


Figure 4.10: Multiple Peaks

Finally, for different categories of entities, limiting the number of years before the peak found to analyze is necessary. The more years observed, the more likely to get a false result. For example for IT products ,there is no need to check 10 years before the peak, because there is a very low chance that a product was mentioned that long ago before its release date.

The algorithm of year estimation is described as follows:

```
int stop=0.1*MAX_HEIGHT;
int continue=0.8*MAX_HEIGHT;
int defaultYearsToObserve=8;
int[] years={1990,...,2011};
int[] bars={height1,...,heightn};
//find the highest bar bars[n] followed by other bars
int estimate(int index,int times){
    if(times==0||n==0||bar[n-1]==0)
        return year[n];
    else if(bars[n-1]<stop){
        if(bars[n]>continue)
            return n-1;
        else
            return n;
    }else
        return estimate(index-1,times-1);
}
```

Code Listing 4.1: Year Estimation Algorithm

4.3.2 Date Selection

After estimating the first appearance year, it is time to select a possible date from that year. FirstApp Finder retrieves the dates from the timeline's result set of links listed below the timeline chart. However, because the possible interferences of irrelevant dates, the dates listed are double-checked using Palladian toolkit's webpage age detector. Palladian use several methods to detect the age of a webpage: via the URL of a page; via HTTP header of a page; via dates in the structure and content of the web page; and via inbound links from archives[DU11]. It assigns different rates to the dates recognized from the page and returns the best rated date as the result. The precision of Palladian Webpage aget detector is measured above 70%. The date selection algorithm is illustrated below:

```
int year=ESTIMATED_YEAR;
Date[] tlDates={d1,...,dn};
Date[] pgAges={a1,...,an};
Date youngest=min(pgAges);
if(youngest.YEAR!=year)
    return d1;
else
    return youngest;
```

Code Listing 4.2: Date Selection Algorithm

Chapter 5

Evaluation

This section presents an evaluation of the effectiveness of FirstApp Finder.

5.1 Sample Dataset Size

We are aiming at a confidence interval of 95% confidence level with a 10% variance. According the normal distribution table and the following formula:

$$n = \frac{z^2\sigma^2}{d^2}(\sigma = 0.5, z = 1.96, d = 10\%) \tag{5.1}$$

the number of samples is 96. In the evaluation, 100 entities are included. All entities are listed in table 5.1. The column Release Time is when the entity actually available. For actors, it is when they go in public or have official works. The column First Appearance is the earliest date found on the Web. Note that because of the lack of relevant tools, the first appearance dates are manually searched as early as possible. The keywords used for the search are "entity name" plus category name. For example, "My Big Fat Greek Wedding" as a whole word and the word "movie".

Category	Name	Release Time	First Appearance	Estimated Year	Estimated Date
Movie	My Big Fat Greek Wedding	2002	25,Jan,2001	2001	Jan 25 2001
Movie	My Big Fat Greek Wedding	2002	25,Jan,2001	2001	Jan 25 2001
Movie	Million Dollar Baby	2004	25,Feb,2004	2004	Feb 25 2004
Movie	Ocean's Twelve	2004	23,Sep,2002	2002	Sep 23 2002
Movie	Slumdog Millionaire	2008	11,May,2007	2007	May 11 2007
Movie	Pirates of the Caribbean: Dead Man's Chest	2006	03,Feb,2005	2005	Feb 03 2005
Movie	Ratatouille	2007	07,Aug,2004	2004	Feb 29 2004
Movie	Blood Diamond	2006	16,Mar,2005	1993	Mar 16 2005
Movie	Transformers: Revenge of the Fallen	2009	04,Oct,2007	2007	Oct 04 2007
Movie	The Spy Next Door	2010	08,Jul,2008	2007	Jul 15 2007

Movie	Piranha 3DD	2011	19,Aug,2010	2010	Aug 19 2010
Mobile phone	Nokia 3510i	2002	06,Sep,2002	2002	Sep 06 2002
Mobile phone	LG VX4400	2003	04,Feb,2003	2003	Feb 04 2003
Mobile phone	BenQ P30	Oct,2004	13,Mar,2003	2003	Mar 13 2003
Mobile phone	Samsung SGH-D307	2005	19,May,2005	2004	Jun 15 2005
Mobile phone	O2 Xda Neo	Mar, 2006	24,Jan,2006	2006	Jan 24 2006
Mobile phone	Apple iphone	29,Jun,2007	14,Feb,2005	2006	Jan 17 2006
Mobile phone	Nokia N95	2007	26,Sep,2006	2006	Sep 21 2006
Mobile phone	Motorola A455	2009	04,Mar,2009	2009	Mar 04 2009
Mobile phone	Nokia E7-00	2011	10,Jul,2009	2009	Jul 10 2009
Mobile phone	Nokia 5230	Nov, 2009	28,Apr,2008	2008	Jan 04 2008
PC game	Age of Wonders 2	2002	27,Jan,2001	2001	Jan 27 2001
PC game	Delta Force Black Hawk Down	2003	10,Jul,2002	2002	Aug 02 2002
PC game	Counter-Strike: Source	11,Aug, 2004	01,Jan,2004	2004	Jan 01 2004
PC game	World of Warcraft	23,Nov, 2004	01,Sep,2001	2002	Aug 12 2002
PC game	UFO: Aftershock	2005	15,Jun,2004	2004	Jun 15 2004
PC game	Cloning Clyde	19,Jul, 2006	09,May,2006	2006	May 09 2006
PC game	BioShock	2007	09,Jan,2006	2006	Jan 10 2006
PC game	Plants vs. Zombies	2009	31,Jan,2008	2008	Jan 31 2008
PC game	NBA 2K11	October5, 2010	11,Oct,2009	2009	Oct 11 2009
PC game	Magicka	2011	06,Oct,2009	2009	Oct 06 2009
Graphics card	GeForce4 MX 440	2002	18,Mar,2002	2002	Mar 18 2002
Graphics card	GeForce FX5900 ULtra	May 2003	12,Jan,2003	2003	Jan 12 2003
Graphics card	GeForce 6600	August 12, 2004	09,Jul,2004	2004	Jul 09 2004
Graphics card	Radeon X1300	2005	30,Sep,2005	2004	Jan 01 2004
Graphics card	Radeon X1950 XTX	August 23, 2006	11,Mar,2006	2006	Jan 24 2006

Graphics card	GeForce 8300 gs	July,2007	08,Apr,2007	2007	Apr 17 2007
Graphics card	Radeon HD 4850	June 19, 2008	07,Jan,2008	2008	Jun 19 2008
Graphics card	GeForce GT 240	17,Nov 2009	12,Oct,2009	2009	Nov 01 2009
Graphics card	GeForce GTX 570	7?December 2010	12,Jul,2010	2010	Jul 12 2010
Graphics card	GeForce GT 440	1,Feb 2011	17,Oct,2010	2010	Oct 17 2010
Mp3 player	Ipod shuffle	Jan 11, 2005	19,Jan,2004	2004	Jan 19 2004
Mp3 player	Archos Jukebox Multimedia	2002	14,Dec,2001	2001	Dec 14 2001
Mp3 player	NOMAD Jukebox Zen NX	August 20, 2003	23,Jul,2003	2003	Jul 23 2003
Mp3 player	Rio Karma	Aug 2003	11,Aug,2003	2003	Aug 11 2003
Mp3 player	Cowon iAUDIO U2	July 2004	20,Sep,2005	2005	Sep 20 2005
Mp3 player	Creative ZEN Sleek	30 ,08, 2005	07,Jun,2005	2005	Jun 07 2005
Mp3 player	Samsung YP-S5	30,08, 2007	16,Aug,2007	2007	Aug 16 2007
Mp3 player	Sony NW-A919	November 2007	27,Sep,2007	2007	Sep 27 2007
Mp3 player	Philips SA075	29,12,2009	29,Dec,2009	2009	Dec 29 2009
Mp3 player	Samsung Galaxy Player 50	2011	03,Sep,2010	2010	Dec 31 2010
Laptop	Apple iBook G3	May 1, 2001	25,Jan,2000	2001	Dec 31 2001
Laptop	Dell Inspiron 9300	Feb 24, 2005	12,Jan,2005	2004	May 19 2004
Laptop	Apple PowerBook G4	Jan 2001	08,Jan,2001	2000	Jan 01 2000
Laptop	Fujitsu LifeBook S710	January 22, 2010	02,Jun,2010	2010	Jul 27 2010
Laptop	Asus Eee PC 4G	October 16, 2007	01,Nov,2007	2007	Nov 01 2007
Laptop	Acer Aspire One	July, 2008	12,Feb,2008	2008	May 29 2008
Laptop	Apple MacBook Air	15,01,2008	31,Mar,2007	2007	Mar 31 2007
Laptop	Dell Adamo 13	March 17, 2009	18,Mar,2009	2008	Dec 31 2008
Laptop	HP Envy 13	October 15, 2009	12,Sep,2009	2009	Sep 12 2009
Laptop	iPad	April 3, 2010	05,May,2009	2009	Jan 21 2009
Song Sheryl Crow	Soak up the sun	2002	19,Apr,2002	2002	Apr 15 2002

Song U2	Sometimes You Cant Make It on Your Own	7 February 2005	02,Sep,2004	2004	Sep 02 2004
Song Alicia Keys	If I Ain't Got You	February 17, 2004	07,Aug,2003	2002	Dec 31 2002
Song Herbie Hancock	River: The Joni Letters	September 25, 2007	06,Dec,2007	2007	Dec 06 2007
Song Fergie	Big Girls Dont Cry	May 15, 2007	18,Sep,2006	2005	Dec 31 2005
Song Kings of Leon	Sex on Fire	5, Sep, 2008	28,Jun,2008	2008	Aug 05 2008
Song Metallica	My Apocalypse	26, Aug , 2008	18,Jul,2008	2008	Jul 31 2008
Song Dave Matthews Band	Big Whiskey and the GrooGrux King	2, Jun, 2009	22,May,2009	2008	Dec 31 2008
Song pink	Glitter in the Air	31, Jan, 2010	24,Oct,2008	2008	Oct 24 2008
Song lady gaga	Born This Way	11,Feb, 2011	22,Mar,2010	2010	Mar 22 2010
Digital camera	Kyocera Finecam S3R	2,Dec, 2003	07,Jul,2004	2003	Dec 31 2003
Digital camera	Olympus C-750 Ultra Zoom	19,Jun, 2003	10,Mar,2003	2003	Apr 15 2003
Digital camera	Canon PowerShot S500	06,May,2004	09,Jan,2004	2004	Jan 09 2004
Digital camera	canon Ixus 430	09,Feb,2004	07,Jun,2004	2004	Oct 20 2004
Digital camera	Panasonic Lumix DMC-L10	Sep,2007	30,Aug, 2007	2007	Jun 27 2007
Digital camera	Nikon D700	01,Jul,2008	09,May,2008	2007	Dec 31 2007
Digital camera	Olympus PEN E-P1	16,Jun, 2009	13,Jul,2009	2008	Dec 31 2008
Digital camera	Nikon Coolpix L22	3,Feb , 2010	05,Feb,2010	2010	Feb 1 2010
Digital camera	Canon G12	1, October , 2010	05,May,2010	2009	Dec 31 2009
Digital camera	Nikon Coolpix S5	21, Feb, 2006	19,Dec,2005	2005	Dec 19 2005
Company	BitPass	2002	12,Jan,2003	2003	Jan 12 2003
Company	IDology	2003	29,Jun,2003	2003	Jul 26 2003
Company	Diskoline	2004	20,Sep,2006	2006	Dec 31 2006
Company	Energy Alberta Corporation	2005	01,Mar,2007	2007	Mar 12 2007

Company	inniAccounts	2006	05,May,2010	2010	May 05 2010
Company	IMINT Image Intelligence AB	2007	08,Jan,2009	2009	Dec 31 2009
Company	TrustPort	2008	09,Mar,2006	2005	Dec 31 2005
Company	Wind Mobile	2009	11,Feb,2008	2007	Jan 30 2007
Company	Verba Technologies	2010	20,Jun,2010	2010	Jun 29 2010
Company	Fokker Technologies	01, Jan, 2011	18,Nov,2010	2010	Dec 31 2010
Actor	Emma Watson	2001	21,August,2000	2002	Apr 11 2002
Actor	Abigail Breslin	2002	12,May,2002	2004	May 07 2004
Actor	Rhiannon Leigh Wryn	2003	05,Jan,2007	2007	Jan 10 2007
Actor	Chace Crawford	2005	30,Aug,2006	2006	Sep 05 2007
Actor	Nina Dobrev	2006	14,Nov,2006	2007	Nov 14 2006
Actor	Steven R. McQueen	2005	18,Sep,2005	2008	Sep 18 2005
Actor	Dakota Blue Richards	2007	31,Jul,2006	2006	Jul 31 2006
Actor	Sara Canning	2008	11,Jan,2008	2008	Nov 02 2008
Actor	Isabella Acres	2006	28,Jan,2009	2009	Jan 28 2009
Actor	Mia Talerico	2010	08,Mar,2010	2010	

Table 5.1: 100 Entities

5.2 Experimental Result

The result shows the year estimation algorithm of FirstApp Finder is of 80% precision (estimated year matches in actual first appearance year). The full date returned by FirstApp Finder is of 51% precision (Figure 5.1), i.e. totally matched the manually found first appearance date on the Web of an entity. Other dates are averagely 85 days off the first appearance date. The precisions and variance of every category are shown in Figure 5.2 and 5.3.

5.3 Conclusion of Evaluation

From the above results, we conclude that FirstApp Finder has a high average precision of 80% when estimating the year of first appearance of an news entity. It can identify the correct full date of the first appearance of an entity by 51% precision as well (Figure 5.1). For some categories of entities, such as MP3 players, PC games, and movies, it reaches both high precisions in the meantime (Figure 5.2). In most cases, the variance is within 3 months (Figure 5.3). However, it does not quite suit entities such as name of a person or a company, since there might be many articles mentioned other entities with the same name, or there are many biographical articles introducing that named entity, thus many dates appeared in the articles which interferes the estimation of the year of the first appearance of that name, e.g. the high variance of the category ”Laptop” is mainly because of a large number of articles full of irrelevant dates introducing the history of Apple. Once the year of first appearance is estimated incorrectly, the variance between the actual first appearance date and the results returned by FirstApp Finder increases dramatically, as shown in Figure 5.3.

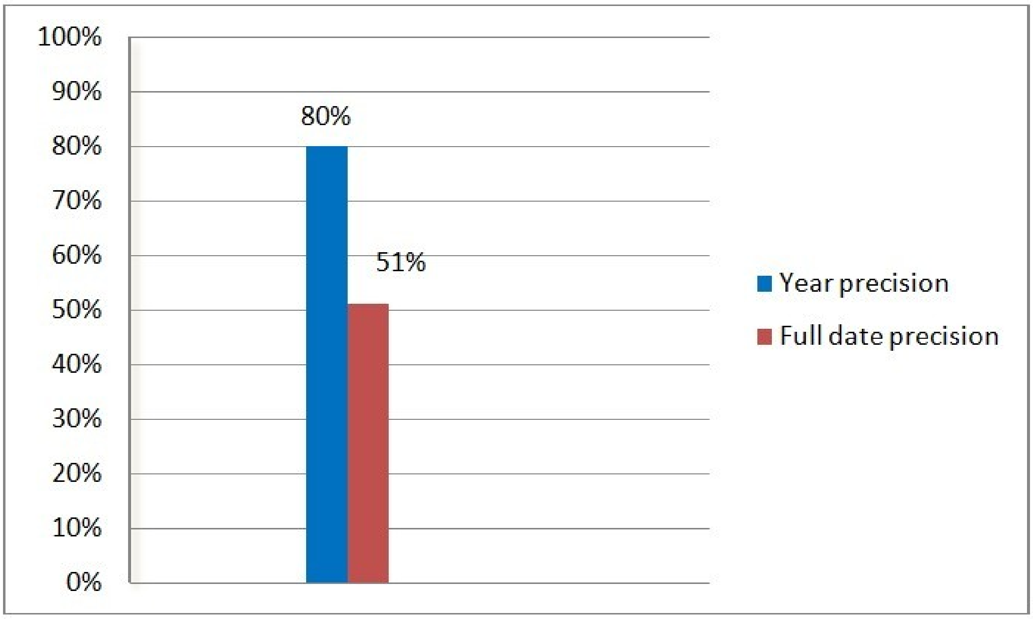


Figure 5.1: Total Precisions

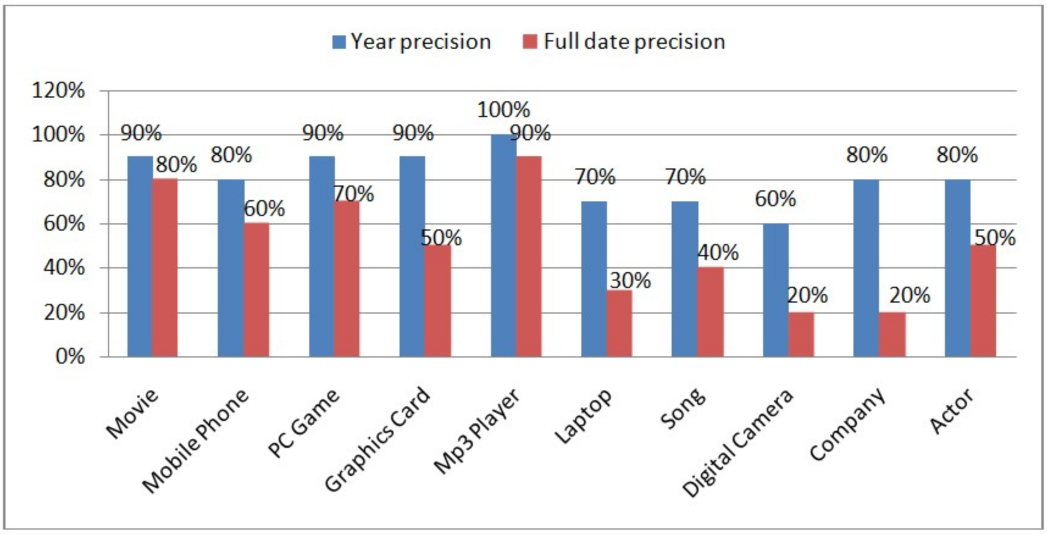


Figure 5.2: Precisions of Separate Categories

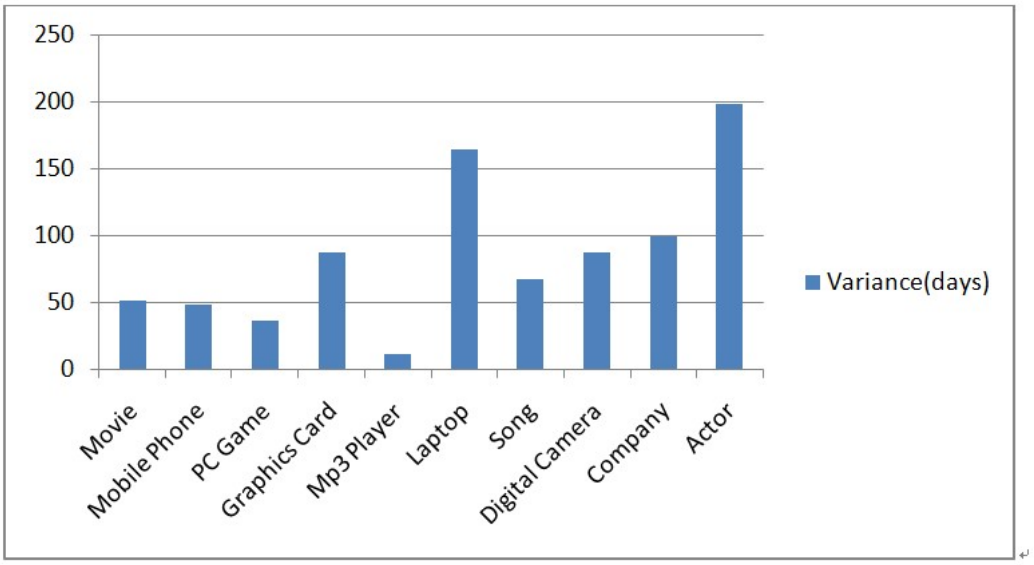


Figure 5.3: Variance of Separate Categories

Chapter 6

Conclusion and Future Work

In this article, a seldom touched question: ”how to find the first appearance of a news entity” is researched. A possible approach based on Google’s timeline operation is proposed and a prototype, FirstApp Finder, is implemented as well. Evaluation of the prototype shows that this approach could be used to estimate the first appearance year and thus find the first appearance date of an entity.

Although FirstApp Finder has a good precision, it is still a long way from being reliable. The year estimation algorithm can be improved to be more intelligent. It is desirable that the program could self-adjust to make new estimations when it finds that the current estimation tends to be wrong. It is not easy since there is no standards to automatically tell when an estimation is likely to be incorrect, except that the program fails to find any possible date in a year. However, this situation is not common.

Besides, the age detector of webpages in Palladian toolkit needs to be improved. Currently it is not completely reliable and returns many false results. On the other hand, FirstApp Finder relies on it very much to judge if a date is truly relevant to the entity being searched. It will be difficult for FirstApp Finder to find the correct date when the age detector fails to provide a right answer. However, without clear semantics in the content of webpages, it is almost impossible to achieve a 100% precision when detecting page ages.

In addition, mechanisms to distinguish adds from normal page contents, registration dates of users from post dates of posts and so on should be developed as well in order to get a higher precision. Finding more websites that lead to false results and puting them into blacklist could help raise the precision as well.

Moreover, a large dataset with numerous entities could be involved in order to find a more precise stopping point of the year estimation.

Bibliography

[BC09] Michael Bendersky and W. Bruce Croft. Finding text reuse on the web. February 2009. 3

[DU08] Marius Feldmann David Urbansky, James A. Thom. Webknox: Web knowledge extraction. December 2008. 1

[DU11] Philipp Katz David Urbansky, Klemens Muthmann. Tud palladian overview. page 5, Febuary 2011. 2.3, 4.3.2

[JAK01] R. Gupta J. Allan and V. Khandelwal. Temporal summaries of news topics. 2001. 3

[SJ00] R. Swan and D. Jensen. Timemines: Constructing timelines with statistical models of word usage. 2000. 3