



# **Mining and Recommending Software Features across Multiple Web Repositories**

Yu Yue

Trustie Group of NUDT

## Introduction

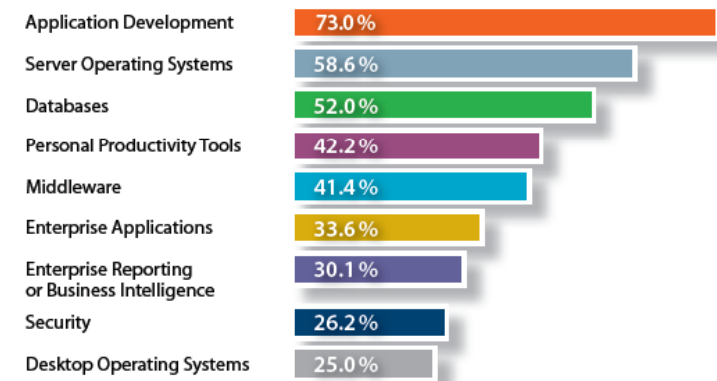
**Begin with an example**

# Introduction

**1. Massive number of OSS projects in open source communities**

**2. OSS projects are utilized for industry wildly**

Community	Time	Scale
Freecode	1998	45,021+
SourceForge	1999	432,004+
Rubyforge	2003	9,349+
Ohloh	2004	500,000+
Google Code	2006	250,000+
...		



[Actuate09] 4<sup>th</sup> Actuate Annual Open Source Survey

## Introduction

**Software Resource**  
**Reorganized in a efficient way**

# Introduction

AIX AJAX Alpha Android Apache 2.0 API Application Frameworks Arcade Archiving Artificial Intelligence backup bash  
Beta Browsers BSD BSD Revised Build Tools C C# C++ Capture CGI Tools/Libraries Chat Clustering/Distributed  
Networks Code cleanup Code Generators **Communications** Compilers Conversion Cryptography CSS Cygwin  
**Database** Database Engines/Servers Debian **Desktop Environment** Development Documentation Dutch **Dynamic**  
**Content** Editors education Email English File Sharing Filesystems Filters Financial Firewalls FreeBSD Freeware  
French Front-Ends Games **Games/Entertainment** General German GNOME GPL GPLv3 Graphics groupware GUI  
Hardware HP-UX HTML HTML/XHTML HTTP Servers Indexing/Search Information Management Initial freshmeat  
announcement Installation/Setup **Internet** Interpreters Italian Japanese Java Java Libraries JavaScript jQuery LGPL  
**Libraries** Library Linux Linux Distributions Log Analysis Logging Mac OS X Markup Mathematics **Monitoring** MP3  
MPL **multimedia** MySQL NetBSD **Networking** News/Diary Objective C **Office/Business** OpenBSD Operating System  
Kernels OS Independent Perl Perl Modules PHP php classes Players POSIX Presentation Puzzle Games Python  
Python Modules Quality Assurance Red Hat Ruby Russian Scheduling Scientific/Engineering Security Server  
Shareware Site Management **Software Development** Solaris Sound/Audio Spanish SQL Stable **Systems**  
**Administration** Tcl Telephony Terminals test Testing Text Editors Text Processing tools Ubuntu Unix Unix Shell  
User Interfaces **Utilities** Version Control Video Viewers Visualization **Web** Windows XML

java (23403)

framework (8300)

linux (6186)

windows (3621)

net (3190)

database (2828)

flash (2257)

flex (2120)

rails (2006)

python (13430)

c (7218)

game (6104)

ruby (3603)

plugin (3106)

c++ (2540)

android (2215)

jquery (2119)

html (1979)

php (12717)

web (6975)

mysql (4059)

ajax (3494)

csharp (3100)

perl (2393)

test (2153)

tools (2092)

server (1881)

gme (11513)

javascript (6703)

library (3810)

xml (3412)

cms (3016)

django (2266)

development (2126)

api (2013)

programming (1779)

mysql

Filters Category: Database x

Refine your search

Category Translations License Programming Language Status OS

Freshness Collection

Search Results for "mysql"

Sort By: Relevance

Showing page 1 of 113.



MySQL

The world's most popular open source database.  
5,513 weekly downloads



mysql client with qt front-end

This is small MySQL graphical client written in C++ using Qt 4.1.  
945 weekly downloads



MySQL-Admin

MySQL-Admin is a simple PHP based administration tool for mysql databases. MySQL-Admin is ea...  
94 weekly downloads



pam-mysql

This is a module that allows PAM aware applications to authenticate users through a MySQL datab...  
57 weekly downloads

# Introduction

Table 1: Labels in open source communities

Repository	total projects	unique labels	ratio (#label=0)	ratio (#label=0,1)
SourceForge	298,402	363	39.80%	77.00%
Ohloh	417,344	102,298	61.68%	69.89%
Freecode	43,864	6,432	8.61%	20.60%

## Introduction

**challenge** for stakeholders to  
retrieve the **suitable** one

# Introduction

sourceforge

Google code

OW2

hloh

确实  
/rustie

free(code):

github

## Stakeholders

### What are you seeking for ?



## Challenge

# Mature Software



**large-scale, uncategorized**

There are **39.8% uncategorized** software in SourceForge  
and **61.68%** in Ohloh

## Challenge

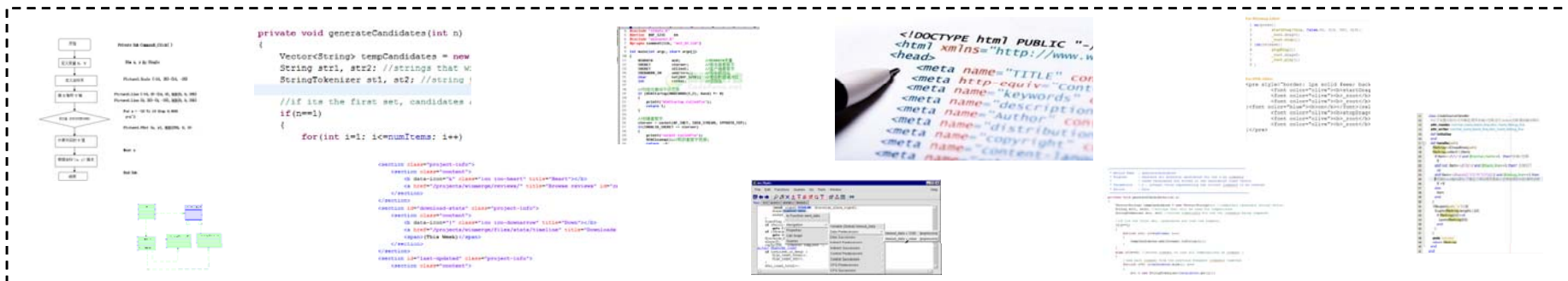
# Reusable Module and Component



**manually , difficultly**

# Challenge

## Reusable Code Snippet



**inaccuracy**

Google Code Search  
Krugle  
Codase Source Code Search Engine



**Feature**

# **Software Feature**

**capture and identify commonalities and differences in  
a software domain**

## **Feature**

**Features are any prominent and distinctive concepts or characteristics that are visible to various stakeholders.**

- a) Features are externally visible characteristics that can differentiate one product from the others.**
- b) Functions, objects, and aspects have been mainly used to specify the internal details of a system.**

**Feature**

**Social Feature**

## Recent releases

[All releases](#)[Release tags](#)

6.6.0 17 Sep 2012 20:57

**Release Notes:** This release adds an Outlook Add-in for Contacts and Calendar exchange syncing between Outlook and Atmail Server using ActiveSync.

6.5.0 17 Sep 2012 20:57

**Release Notes:** This release adds online file storage with sharing and a mobile-optimized UI.

6.3.5 27 Feb 2012 22:50

**Release Notes:** This is a minor release. It corrects uninitialized array usage in the dashboard controller during graph calculation. It corrects an unhandled exception in logsearch when... [\(more\)](#)

## Norman Malware Cleaner description

[+ SHOW PRODUCT DESCRIPTION](#)

Here are some key features of "Norman Malware Cleaner":

- Easy to install and run
- Detect and Remove malware (viruses, Rootkit's, FakeAV, worms and more)
- Utilize advanced Anti-Rootkit technology
- Quarantine module to process the detected files
- Deep scan and cleaning including Norman patented Norman SandBox technology
- Supports Quick- and Deep Scan mode
- New command line function for better tailor scanning across several machines (businesses)
- Daily signature updates available

### # What is GitX?

GitX is a gitk like clone written specifically for OS X Leopard and higher. This means that it has a native interface and tries to integrate with the operating system as good as possible. Examples of this are drag and drop support and QuickLook support.

### # Features

The project is currently still in its starting phases. As time goes on, hopefully more features will be added. Currently GitX supports the following:

- \* History browsing of your repository
- \* See a nicely formatted diff of any revision
- \* Search based on author or revision subject
- \* Look at the complete tree of any revision
- \* Preview any file in the tree in a text view or with QuickLook
- \* Drag and drop files out of the tree view to copy them to your system
- \* Support for all parameters `git rev-list` has

### # License

GitX is licensed under the GPL version 2. For more information, see the attached COPYING file.

### # Downloading

GitX is currently hosted at GitHub. It's project page can be found at <http://github.com/pieter/gitx>. Recent binary releases can be found at <http://pithub.com/pieter/gitx/wikis>.

## Description

Rigs of Rods is a 3D simulator game where you can drive, fly and sail various vehicles using an accurate and unique soft-body physics engine.

[Rigs of Rods Web Site >](#)

### Categories

Simulation, Simulations

### License

GNU General Public License version 3.0 (GPLv3)

## Features

- rigid body physics simulation
- can simulate cars, trucks, airplanes and boats and everything inbetween

## Challenge

# 1. Synonymic Feature Element:

Kills the core of AdPower and not only symptoms;↵

Kills the core of Adroar and not only symptoms;↵

Kills the core of BANCOS.B and not only symptoms;↵

Ability to update that does not require downloading full package;↵

Incremental database updates and often to include information about latest threats;↵



## Challenge

# 2. Hybrid Semantic-level:

- (1) *“Internationalized GUI”;*
- (2) *“Various language packs are available”;*
- (3) *“Multi-language supported: including English, Simplified Chinese, Traditional Chinese, Japanese, Korean, German, French, Spanish, Italian, Russian etc”;*

## Topics

gene	0.04
dna	0.02
genetic	0.01
...	

life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	

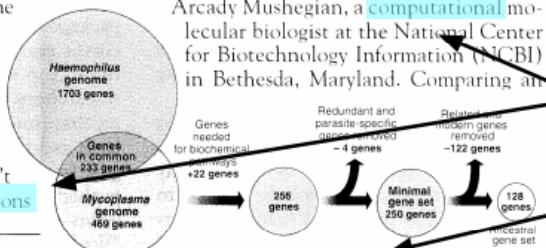
## Documents

### Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many **genes** does an **organism** need to **survive**? Last week at the genome meeting here,\* two genome researchers with radically different approaches presented complementary views of the basic genes needed for **life**. One research team, using **computer** analyses to compare known **genomes**, concluded that today's **organisms** can be sustained with just 250 genes, and that the earliest life forms required a mere 128 **genes**. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those **predictions**

"are not all that far apart," especially in comparison to the 75,000 **genes** in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a **genetic numbers game**, particularly as more and more **genomes** are completely mapped and sequenced. "It may be a way of organizing any newly **sequenced genome**," explains Arcady Mushegian, a **computational** molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



\* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

**Stripping down.** Computer analysis yields an estimate of the minimum modern and ancient genomes.

MAY 1996

## Topic proportions and assignments

Doc – Topic

$\theta$

Topic – Word

$\beta$

# Topic Model

$\theta$

```
1 tion ...
2 \stem\1.txt 18 0.3684210479259491 7 0.21052631735801697 19 0.15789473056793213 2 0.105263
3 \stem\10.txt 22 0.7272727489471436 23 0.1818181872367859 11 0.09090909361839294
4 \stem\100.txt 2 0.3611111044883728 21 0.222222238779068 12 0.1388888955116272 4 0.111111
5 \stem\1000.txt 18 0.6000000238418579 12 0.20000000298023224 13 0.20000000298023224
6 \stem\1001.txt 21 0.3529411852359772 6 0.23529411852359772 2 0.11764705926179886 7 0.1176
7 \stem\1002.txt 19 0.31578946113586426 18 0.21052631735801697 20 0.10526315867900848 21 0.
8 \stem\1003.txt 1 0.1538461595773697 0 0.07692307978868484 5 0.07692307978868484 6 0.07692
9 \stem\1004.txt 11 0.3333333432674408 2 0.13333334028720856 18 0.13333334028720856 7 0.066
10 \stem\1005.txt 0 0.2857142984867096 2 0.1428571492433548 9 0.1428571492433548 19 0.107142
11 \stem\1006.txt 8 0.4000000059604645 1 0.20000000298023224 9 0.20000000298023224 12 0.2000
12 \stem\1007.txt 8 0.375 2 0.125 14 0.125 17 0.125 19 0.125 23 0.125
13 \stem\1008.txt 21 0.2142857164144516 12 0.1428571492433548 14 0.1428571492433548 7 0.107
14 \stem\1009.txt 7 0.1818181872367859 15 0.1818181872367859 18 0.1818181872367859 4 0.0909
15 \stem\101.txt 23 0.29411765933036804 20 0.23529411852359772 0 0.05882352963089943 1 0.05
16 \stem\1010.txt 14 0.5 12 0.25 20 0.25
17 \stem\1011.txt 0 0.3333333432674408 12 0.3333333432674408 6 0.1666666716337204 20 0.1666
18 \stem\1012.txt 2 0.222222238779068 9 0.111111119389534 13 0.111111119389534 15 0.1111
19 \stem\1013.txt 6 0.1818181872367859 7 0.1818181872367859 0 0.09090909361839294 2 0.09090
20 \stem\1014.txt 2 0.4444444477558136 7 0.3333333432674408 11 0.111111119389534 18 0.1111.
```

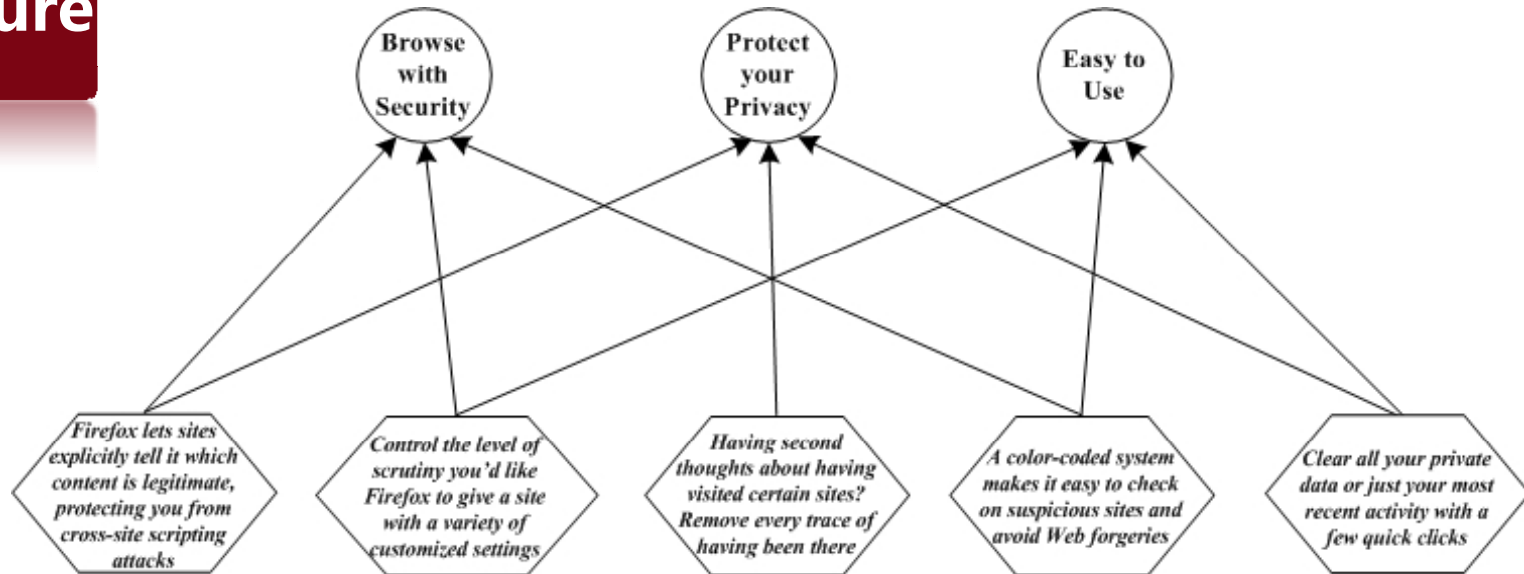
**ID1000: Real-time windows registry monitor utility.**

$\beta$

Topic 12  
run 0.09616858237547893  
window 0.08773946360153256  
includ 0.0739463601532567  
execut 0.04367816091954023  
tool 0.04061302681992337  
microsoft 0.038314176245210725  
task 0.037547892720306515  
outlook 0.02681992337164751  
script 0.022988505747126436  
util 0.022222222222222223

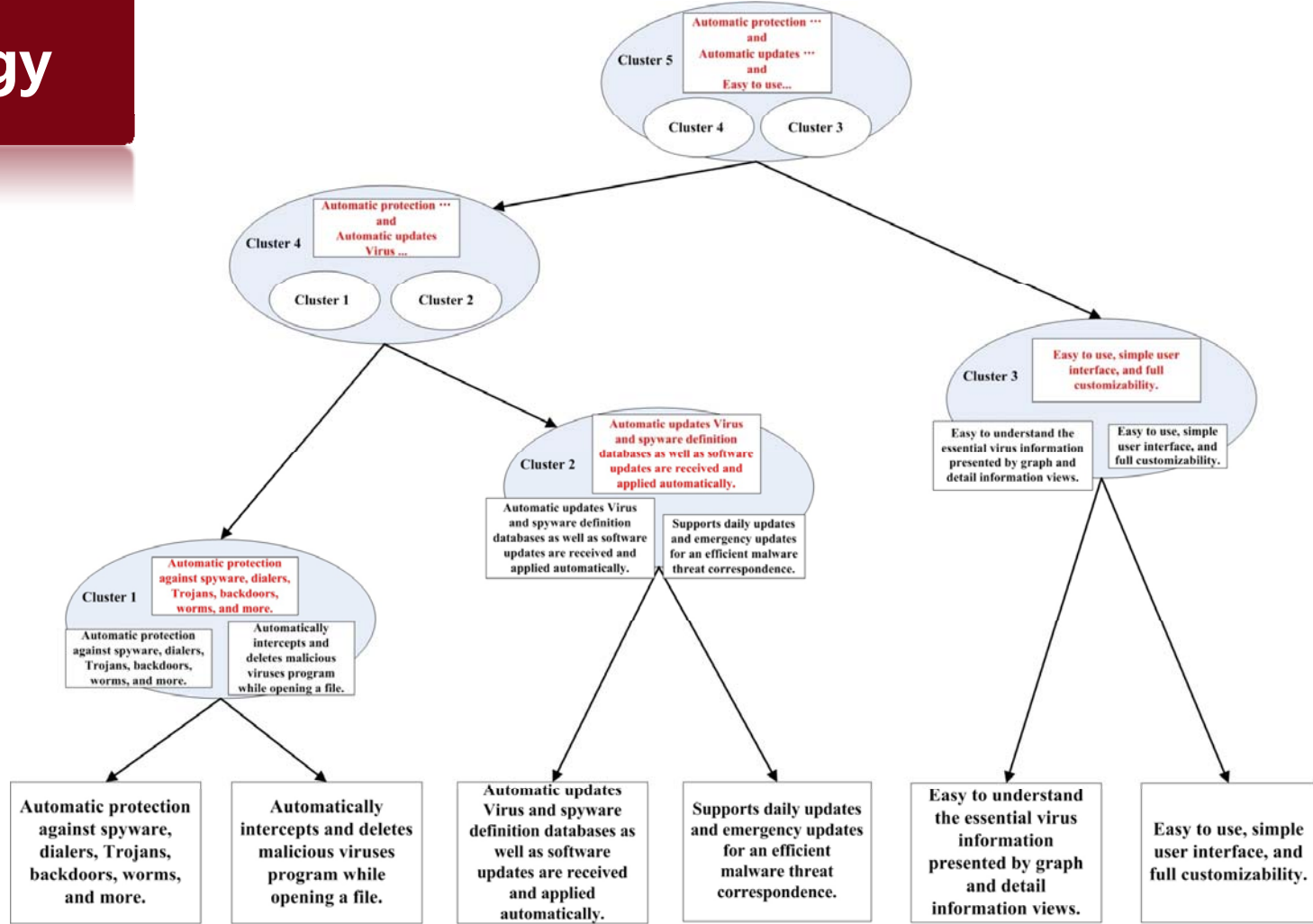
Topic 18  
protect 0.3607470511140236  
threat 0.13990825688073394  
time 0.13663171690694625  
monitor 0.06815203145478375  
real 0.0655307994757536  
dai 0.02162516382699869  
proactiv 0.019003931847968544  
continuu 0.01310615989515072  
hour 0.011795543905635648  
mcafe 0.011795543905635648

# Topic structure

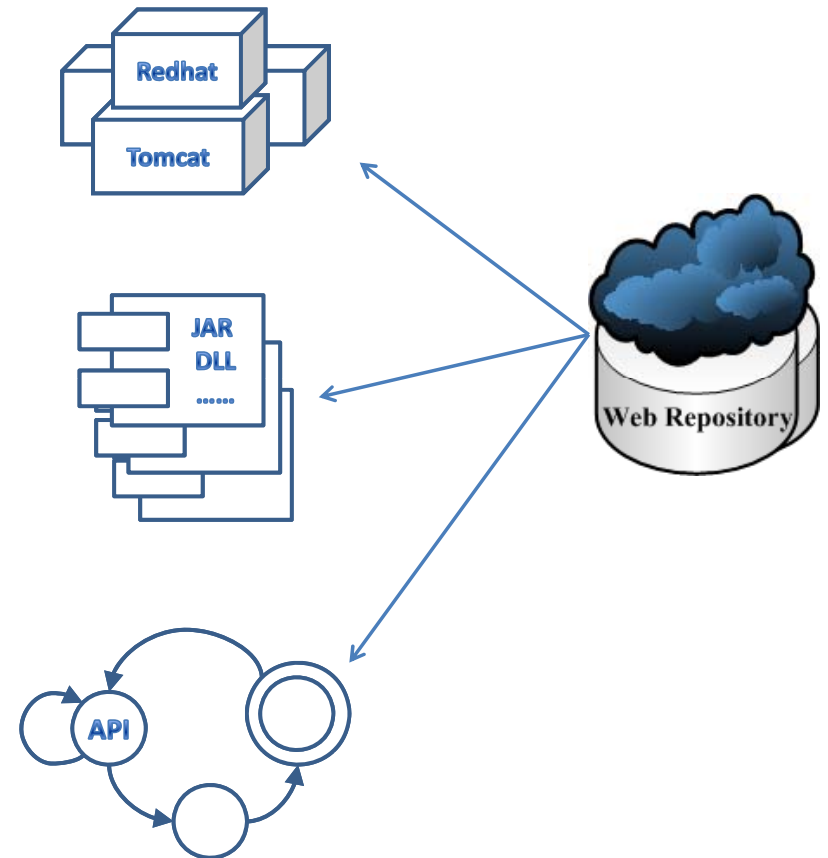
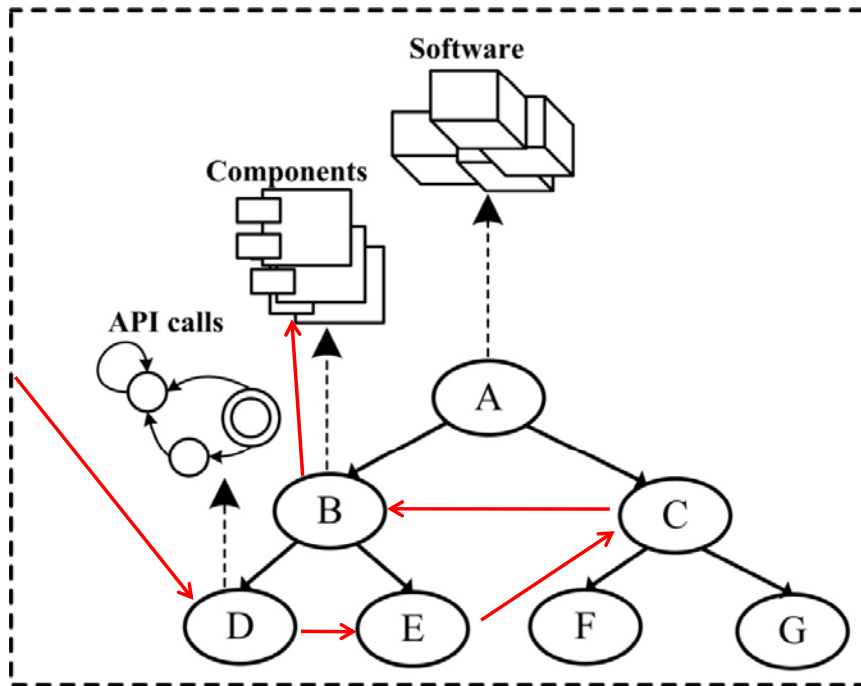
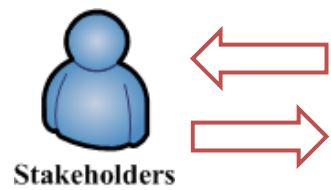


1. Email Scanner enhanced email protection;
2. Email scanning for Microsoft Outlook, Outlook Express, Mozilla Thunderbird, Windows Live Mail, Windows Mail, and other POP3/IMAP mail clients, ensuring your email is free of viruses and other threats;
3. Blocks spam mails, phishing attack mails, junk mails and porn mails before they reach your inbox;

# Ontology



# Overview





# Evaluation

Table 4: Preprocessed experiment datasets

Category	#Feature_sp	#Feature_sf	#Project_sp	#Project_sf	#Topic
Antivirus	2919	1105	667	435	40
Audio-Player	3714	1283	379	530	60
Browser	3010	831	344	177	40
File-Manager	2270	970	330	177	40
Email	8511	1050	823	204	80
Video-Player	3318	2697	379	530	60

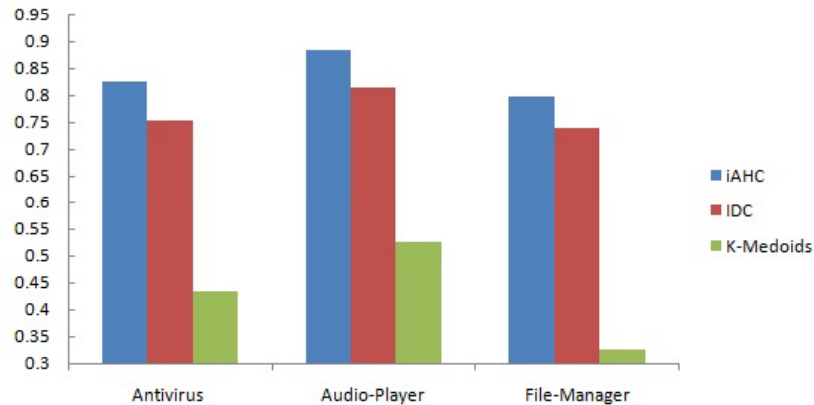
## Evaluation

Table 5: Evaluation of HESA structure

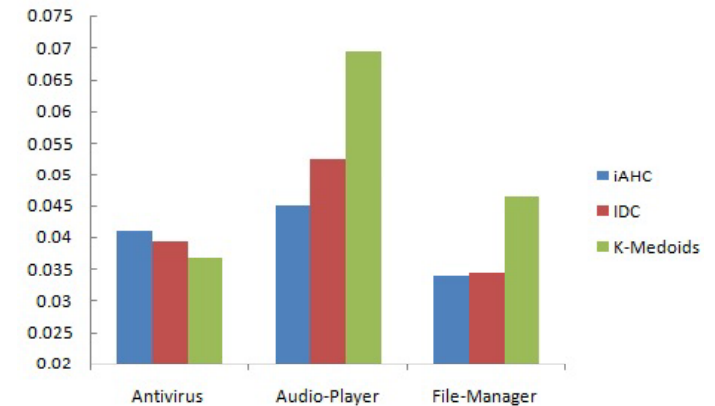
Category	Score-3	Score-2	Score-1	Likert
Antivirus	33.3%	50.0%	16.7%	2.17
Audio-Player	39.1%	46.3%	14.6%	2.25
Browser	36.8%	41.9%	21.3%	2.16
File-Manager	32.7%	52.4%	14.9%	2.18
Email	36.4%	52.8%	10.8%	2.26
Video-Player	47.2%	41.5%	11.3%	2.36
Average	37.58%	47.48%	14.93%	2.23



# Evaluation



(a) the average value



(b) the standard deviation

**The clustering results for each category**

## Recommend

Table 3: An example of resource-by-feature matrix

Software \ Feature	No.21 <sup>1</sup>	No.35 <sup>2</sup>	No.37 <sup>3</sup>	No.41 <sup>4</sup>	No.63 <sup>5</sup>	No.67 <sup>6</sup>
nDVD	0	0	1	1	1	0
FLV Player	1	0	1	1	1	0
Aviosoft DTV Player	1	1	0	0	1	1
Mac Blu-ray Player	1	0	0	1	1	1

<sup>1</sup> Support multi-formats of video files such as FLV, MPEG4, DIVX, HD-MOV, M2TS, MKA, 3GPP and so on.

<sup>2</sup> Smart stretch lets video smart fit on all monitor with different aspect ratio, avoid video loss or distortion.

<sup>3</sup> Title repeat, chapter repeat, AB repeat function that lets you set your favorite scenes for instant repeat.

<sup>4</sup> Fast Forwards and Backwards with Customizable Speeds.

<sup>5</sup> You can easily configure every option of the Player by using a nice preferences dialog.

<sup>6</sup> Brightness, contrast, hue, saturation and gamma settings.

# Recommend

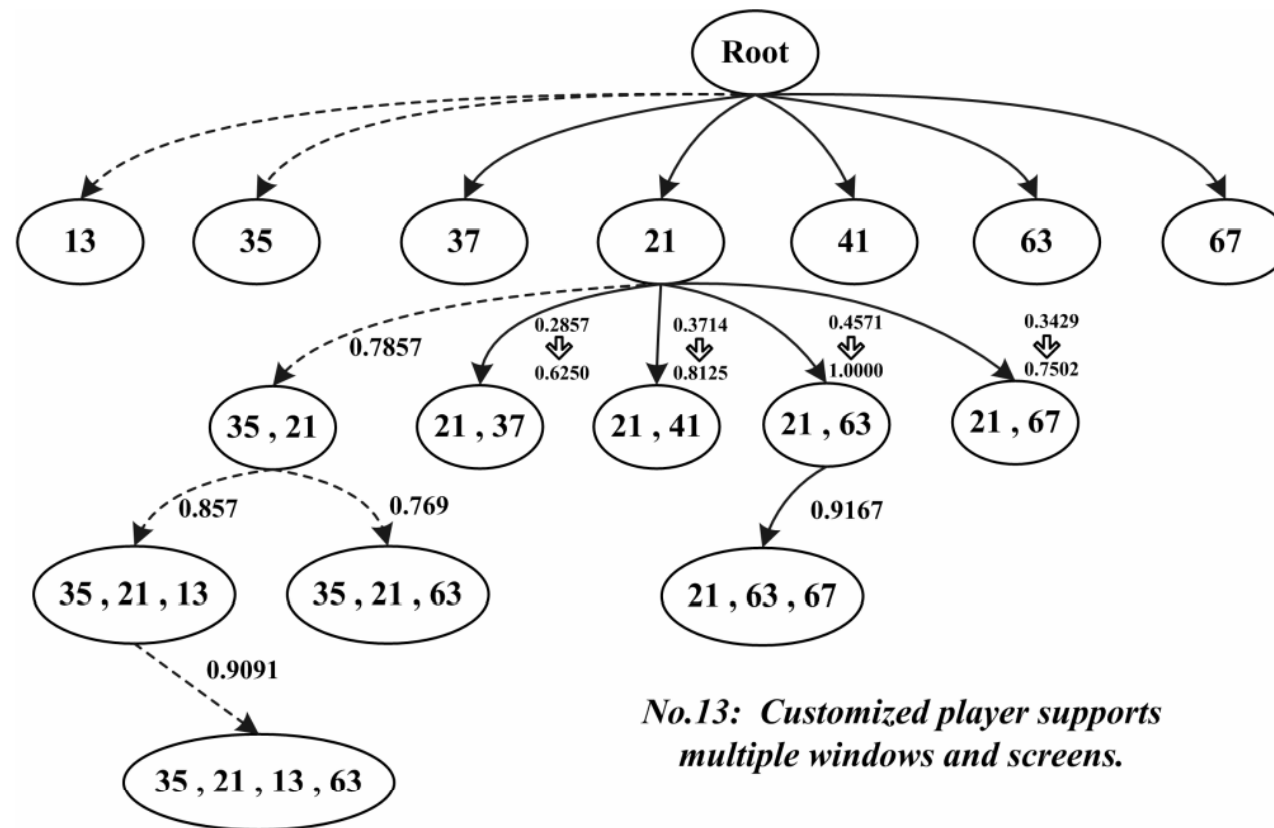


Figure 3: Part of the feature-pattern in Video-Player domain

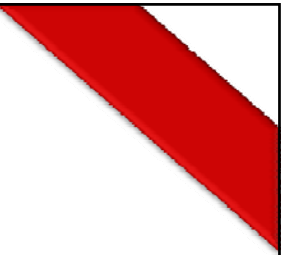
# Recommend

Table 7: Feature recommendations

Category	Input	Recommendations
Email	-High level filter system for spam.	<div>-Automatically start spam process when Windows starts. -Attachment and keyword filtering. -Privacy guaranteed-your emails never leave your network. -Full support for international characters. -Automatic import of local address book.</div>
Video-Player	-Support multi-formats of video.	<div>-Smart stretch lets video smart fit on all monitor with different aspect ratio, avoid video loss or distortion. -Customized player supports multiple windows and screens. -Play anything including movie, video, audio, music and photo. -Video desktop lets you view video in true background mode like wallpaper.</div>
Audio-Player	<div>-Radio streaming. -Easy to use and friendly User Interface.</div>	<div>-Free Lossless Audio Codec. -Playlists for each day of week or date. -Flexible XML based skinning engine, Create your own skins, or choose one of the available skins. -Contextual Help System.</div>

## **Conclusion**

- 1. Build a Hierarchical rEpository of Software feAture (HESA)**
- 2. Induce the Feature-Pattern base on association rule mining**
- 3. Recommend the most relevant features to users**



# Thank You !

Yu Yue  
INFLUX Group  
National University of Defense Technology