



HESA

The Construction and Evaluation of Hierarchical Software Feature Repository

Yu Yue
INFLUX Group
National University of Defense Technology

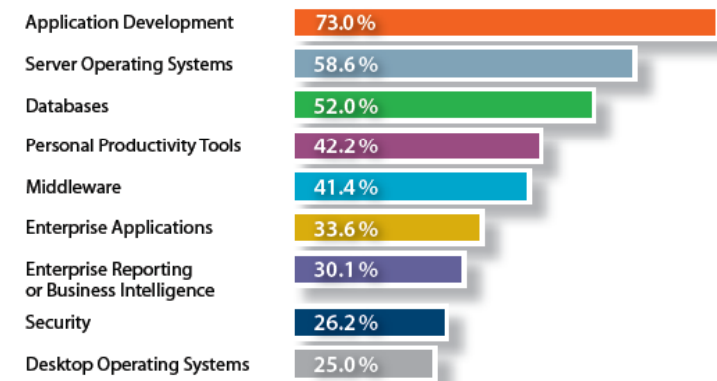
Introduction

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] 4th Actuate Annual Open Source Survey

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

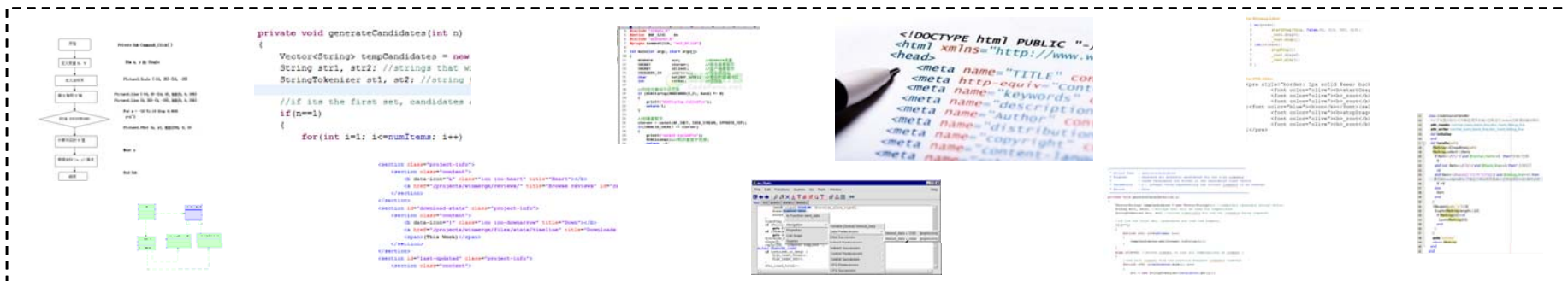
Reusable Module and Component



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:

Email Scanner enhanced email protection;⁺

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;⁺

Blocks spam mails, phishing attack mails, junk mails and porn mails before they reach your inbox;⁺

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
...	

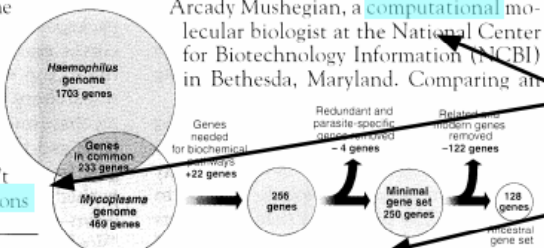
Documents

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

θ

Topic – Word

β

Topic Model

 θ

```
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.
```

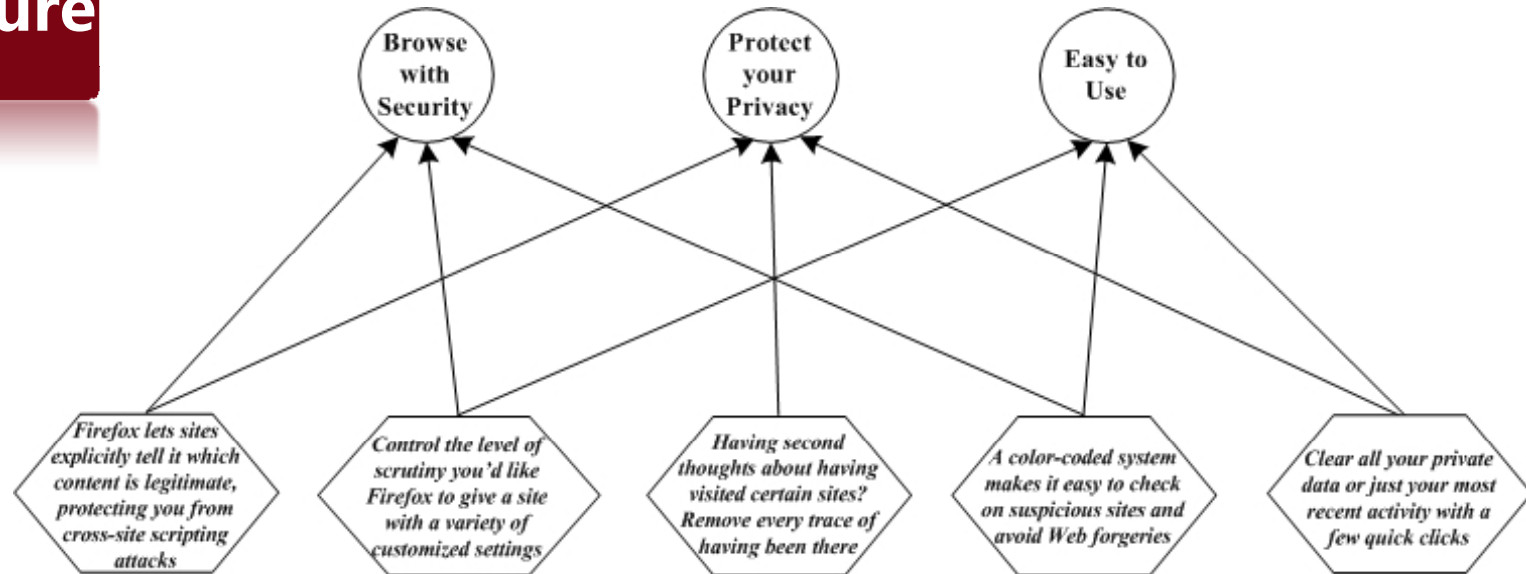
ID1000: Real-time windows registry monitor utility.

 β

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

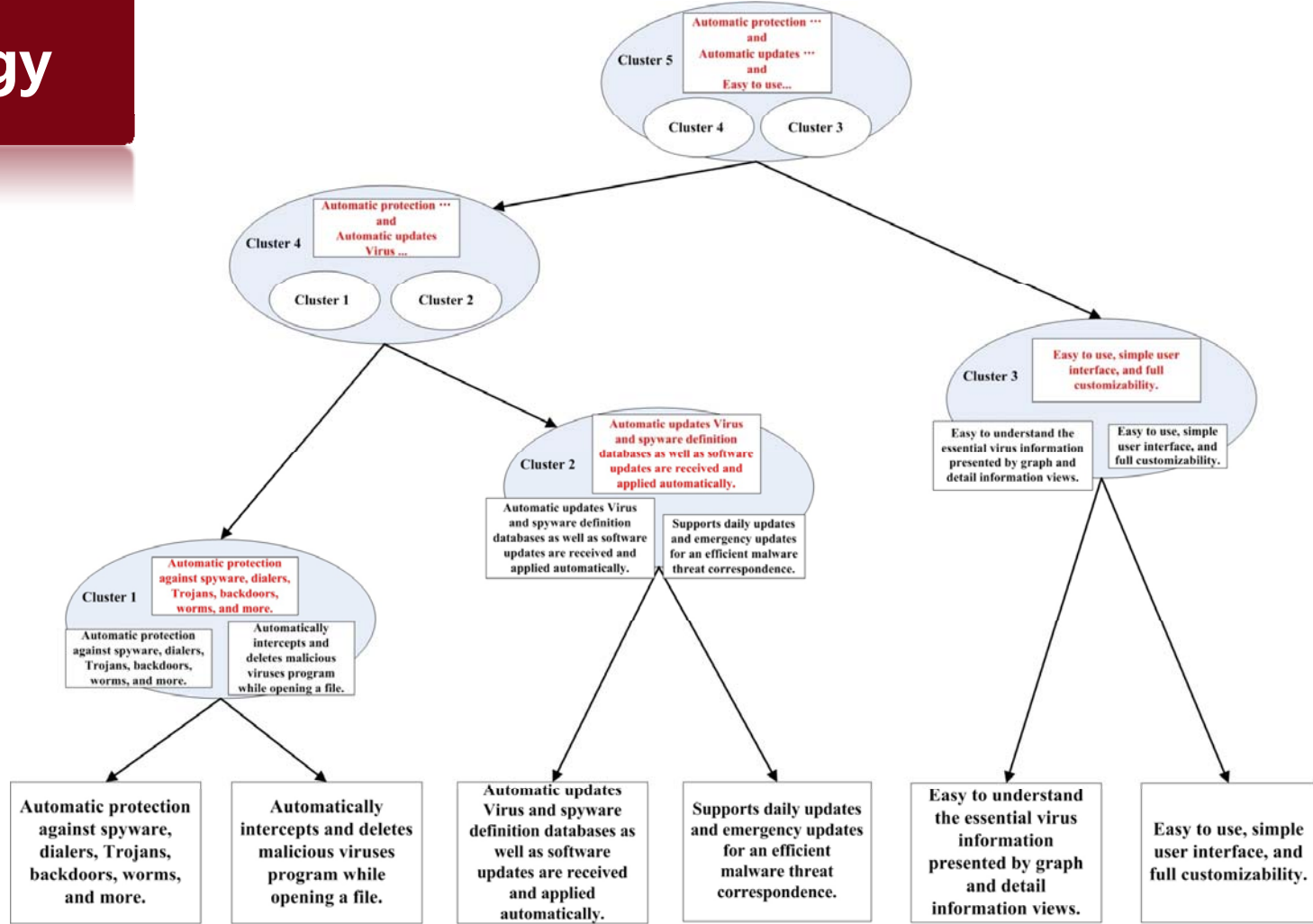
Topic 18
protect 0.3607470511140236
threat 0.13990825688073394
time 0.13663171690694625
monitor 0.06815203145478375
real 0.0655307994757536
dai 0.02162516382699869
proactiv 0.019003931847968544
continuu 0.01310615989515072
hour 0.011795543905635648
mcafe 0.011795543905635648

Topic structure

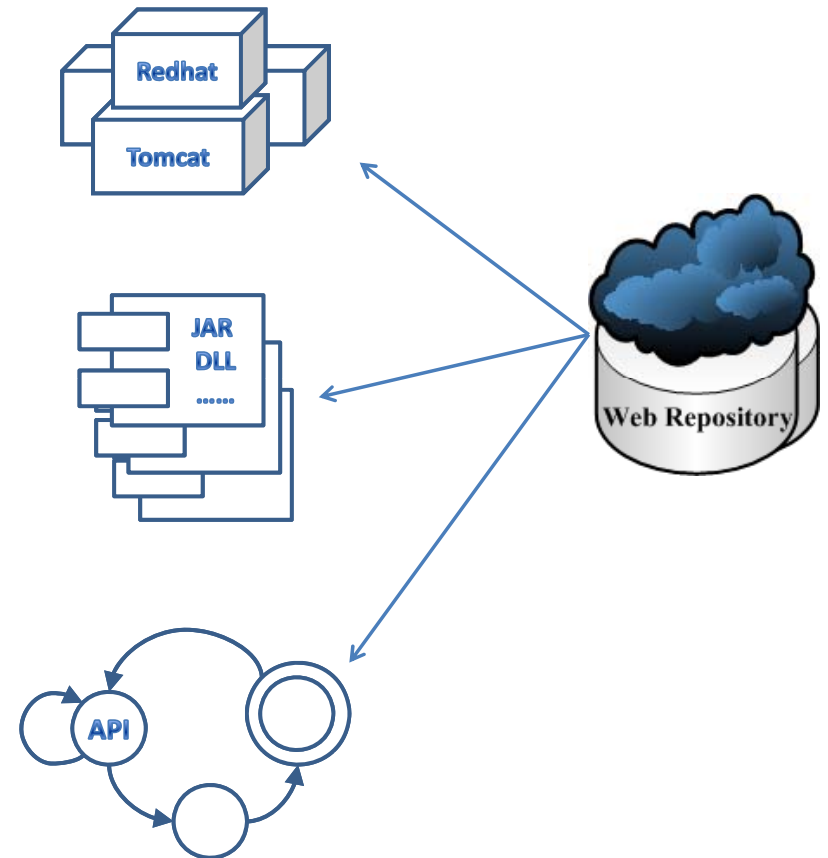
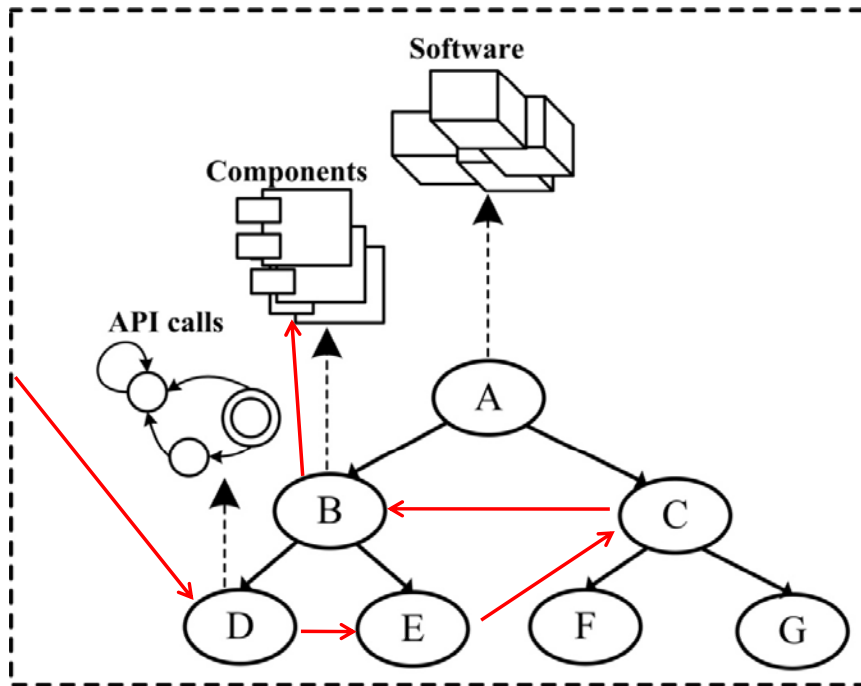
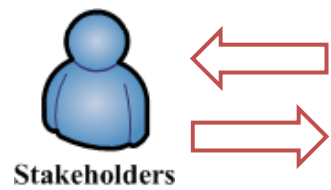


1. Email Scanner enhanced email protection;
2. Email scanning for Microsoft Outlook, Outlook Express, Mozilla Thunderbird, Windows Live Mail, Windows Mail, and other POP3/IMAP mail clients, ensuring your email is free of viruses and other threats;
3. Blocks spam mails, phishing attack mails, junk mails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

Ontology



Overview



Evaluation

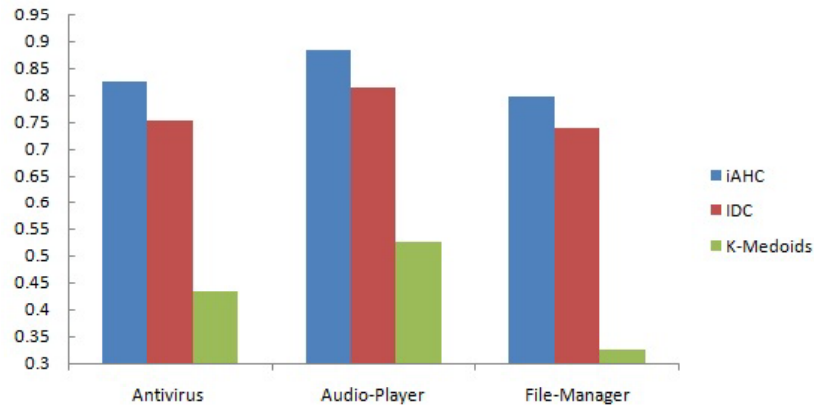
Table I
PREPROCESSED EXPERIMENT DATASETS

Category	#Softpedia	#Sourceforge	#Total	#Topic
Antivirus	2919	1105	4024	50
Audio-Player	3714	1283	4997	60
File-Manager	2270	970	3240	40

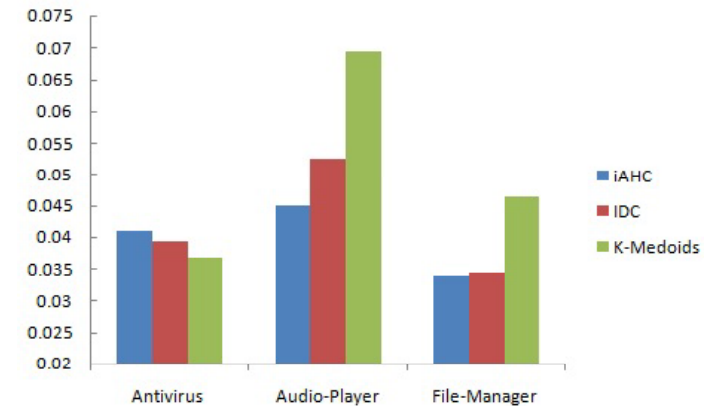
Table II
EVALUATION OF FEATURE-ONTOLOGY QUALITY

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
File-Manager	32.7%	52.4%	14.9%	2.18
Average	35.03%	49.57%	15.4%	2.20

Evaluation



(a) the average value



(b) the standard deviation

The clustering results for each category



Thank You !

Yu Yue
INFLUX Group
National University of Defense Technology