

Intelligent Semantic Web Search Service – The Intute Project

Speaker: Yanbo J. Wang

**NaCTeM, School of Computer Science
University of Manchester**

Project Description

The Intute project, co-funded by JISC (Joint Information Systems Committee) and AHRC (Arts and Humanities Research Council), is a joint work between NaCTeM, Mimas and the Intute Repository Search Project.

The aim of the Intute project is to develop an intelligent semantic web search service using NaCTeM's text mining tools to grant users the benefit of advanced searching within an enhanced subset of the Intute repository, which harvests and aggregates metadata from UK-wide open repositories.

One aspect for the Intute project is to employ the techniques of Text Classification (TC) — automated categorisation of “unseen” documents into pre-defined class-groups.

The Usage of TC in Intute

The “*two-stage*” usage of TC techniques in the Intute project can be detailed as follows.

Stage-one Usage: *Single-label TC*

During the early stages of the Intute project, we are only focusing on those documents belonging to either *Social Science* or Bio-medical Science. However, documents in the Intute repository are not necessarily assigned to domain-classes. It is therefore an essential preliminary task to automatically and accurately distinguish these *Social Science* or Bio-medical Science documents from other documents in the collection.

Stage-one Usage of TC in Intute

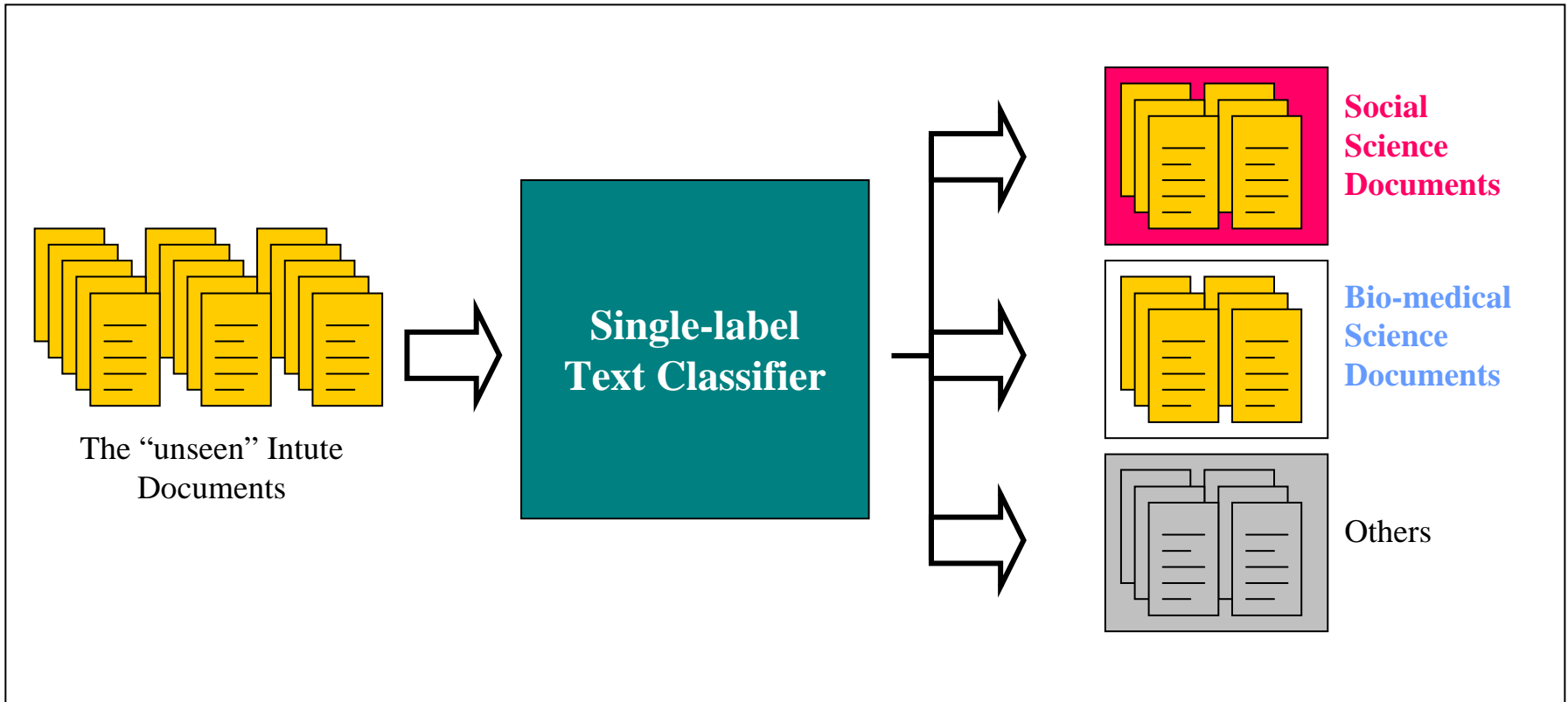


Fig. 1. Stage-one Usage of TC in Intute

Demo of Single-label TC

- The TFPTC text mining software

Classifier Type	CARM – Classification based on Association Rule Mining
Classifier Name	TFPTC – Total From Partial Text Classification
Document-base	Reuters.D6643.C8
# of Documents	6,643
# of Classes	8, {acq, crude, earn, grain, interest, money-fx, ship, trade}
# of Doc. per Class	{2,108, 444, 2,736, 108, 216, 432, 174, 425}
Feature Selection	Mutual Information
# of Key Words	1,200
Support	0.1%
Confidence	35%
Training : Test	50 : 50

```

Command Prompt - java TextMiningGUI_App
C:\Documents and Settings\Yanbo J. Wang\Java Code\PhD_related_software\MI_NGL>ja
vac *.java

C:\Documents and Settings\Yanbo J. Wang\Java Code\PhD_related_software\MI_NGL>ja
va TextMiningGUI_App
*****
*
*          START TEST RUN
*
*****
SETTINGS
Confidence (default 80%) = 35.0
Significance index       = 1.0E-4
Lower noise t'hold      = 0.2
Support (default 20%)  = 0.1
Upper noise t'hold     = 20.0
Maximum Num. Sig. Words = 1200

START ANALYSING DOC BASE FOR KEYWORDS

READ TRAINING SET
=====
Num. docs. in training set = 3321

ID NOISE, ORDINARY AND SIGNIFICANT WORDS IN WORDS BIN TREE
LST = 0.2% (6 docs), UST = 20.0% (664 docs), SI = 1.0E-4
Max # sig words 1200
=====
Significance contribution calculation stratgey = Mutual Information
Potential Sig. Word. list generation stratgey = All Significant Words
Sig. Word. identification stratgey = First N (distributed), N is 1200 words.
Number of one itemsets = 1208
READ TEST SET
=====
Num. docs. in test set = 3322
START CLASSIFICATION (BEST FIRST), TFPC WITH X-CHEKING AND CSA ORDERING
-----
Max num frequent sets = 500000
Max size of antecedent = 6
Number of records in training set = 3321
NOTE: Data set reordered
Creating Postive table
Support = 0.1, Confidence = 35.0
Minimum support = 3.32 (Records)
Max num frequent sets = 500000
Max size of antecedent = 6
Number of records in training set = 3321
NOTE: Data set reordered
Apriori-TPP with X-Checking
Minimum support threshold = 0.1% (3.32 (records)
Generation time = 33.0 seconds (0.55 mins)
Accuracy 85.01
Number of rules 1503
Max size of antecedent = 3
Ave size of antecedent = 1.34
Antecedent size distribution
    Antecedent size 1 = 1044
    Antecedent size 2 = 533
    Antecedent size 3 = 6

CLASSIFIER
-----
<0> {accounted} -> {earn} 100.0%
<0> {advance} -> {earn} 100.0%
<0> {affiliates} -> {earn} 100.0%
<0> {aide} -> {earn} 100.0%
<0> {amsterdam} -> {earn} 100.0%
<0> {aug} -> {earn} 100.0%
<0> {borrowers} -> {earn} 100.0%
<0> {bringings} -> {earn} 100.0%
<0> {carrying} -> {acq} 100.0%
<0> {citing} -> {earn} 100.0%
<0> {comments} -> {trade} 100.0%
<0> {complaints} -> {acq} 100.0%
<0> {cos} -> {acq} 100.0%
<0> {coupon} -> {earn} 100.0%
<0> {deterioration} -> {earn} 100.0%
<0> {distribution} -> {acq} 100.0%
<0> {dollars} -> {earn} 100.0%
    
```

LUCS-KDD: Text Mining GUI

File	Thresholds	Sig. Word Strats.	Algorithm	Evaluation	Class. Strat.	Output	Batch
About							
Text Base Dir.							
Text Base Spec.							
First N recs							
Start							
Start Batch Mode							
Start Set of 8 Mode							
Exit							

OUTPUT THRESHOLD VALUES:
Support = 0.1
Confidence = 35.0
Significance = 1.0E-4
Upper Noise = 20.0
Lower Noise = 0.2

SET MAXIMUM NUMBER SIGNIFICANT WORDS:
Maximum number of permitted significant words = 1200

Selected significance contribution calculation strategy = Mutual Information

Selected potential significance words list generation strategy = All Significant Words

Selected significant word ID strategy = First N (distributed)

Selected TM algorithm = Keywords

Selected rule evaluation strategy = 50:50

Output:
Classifier Stats. = true

Output:
Classifier Stats. = true
Classifier = true

LUCS-KDD (Liverpool University Computer Science - Knowledge Discovery
in Data) group Text Mining demonstrator.

Version 1 Created by Frans Coenen (28 February 2006)

The Keyword-only Approach

```

Command Prompt - java TextMiningGUI_App
(0) <tires> -> <acq> 100.0%
(0) <travel> -> <acq> 100.0%
(0) <treasury> -> <acq> 100.0%
(0) <union> -> <acq> 100.0%
(0) <via> -> <acq> 100.0%
(0) <windfall> -> <acq> 100.0%
(0) <writedowns> -> <acq> 100.0%
(0) <store> -> <crude> 100.0%
(0) <surprised> -> <crude> 100.0%
(0) <trump> -> <crude> 100.0%
(0) <unwillingness> -> <earn> 100.0%
(0) <usair> -> <ship> 100.0%
(0) <taft> -> <trade> 100.0%
(0) <unrelated> -> <trade> 100.0%
(0) <agriculture, crops> -> <interest> 100.0%
(0) <angeles, crops> -> <interest> 100.0%
(0) <banshares, contract> -> <ship> 100.0%
(0) <agricultural, arrived> -> <trade> 100.0%
(0) <brazil, citibank> -> <interest> 100.0%
(0) <agricultural congressmen> -> <interest> 100.0%
(0) <advisors, drilling> -> <interest> 100.0%
(0) <advisors, dependence> -> <money-fx> 100.0%
(0) <banshares, diagnostic> -> <ship> 100.0%
(0) <contract, discount> -> <ship> 100.0%
(0) <banshares, expenses> -> <ship> 100.0%
(0) <declaration, expenses> -> <ship> 100.0%
(0) <banshares, feb> -> <ship> 100.0%
(0) <declaration, feb> -> <ship> 100.0%
(0) <expenses, feb> -> <ship> 100.0%
(0) <casas, calif> -> <acq> 100.0%
(0) <agriculture, foresee> -> <acq> 100.0%
(0) <agriculture, curbs> -> <crude> 100.0%
(0) <banks, damage> -> <grain> 100.0%
(0) <chase, daniel> -> <grain> 100.0%
(0) <daniel, deferred> -> <grain> 100.0%
(0) <daniel, discussed> -> <grain> 100.0%
(0) <deferred, discussed> -> <grain> 100.0%
(0) <deferred, family> -> <grain> 100.0%
(0) <base, grains> -> <grain> 100.0%
(0) <citibank, commodity> -> <interest> 100.0%
(0) <amc, crops> -> <interest> 100.0%
(0) <auctions, crops> -> <interest> 100.0%
(0) <citibank, effective> -> <interest> 100.0%
(0) <advisors, finalized> -> <interest> 100.0%
(0) <bodies, indicates> -> <interest> 100.0%
(0) <bureau, indicates> -> <interest> 100.0%
(0) <agriculture, arab> -> <money-fx> 100.0%
(0) <accord, cereals> -> <money-fx> 100.0%
(0) <advisors, classes> -> <money-fx> 100.0%
(0) <advisors, closes> -> <money-fx> 100.0%
(0) <agriculture, closes> -> <money-fx> 100.0%
(0) <advisors, completing> -> <money-fx> 100.0%
(0) <agriculture, containers> -> <money-fx> 100.0%
(0) <agriculture, dependence> -> <money-fx> 100.0%
(0) <advisors, discussion> -> <money-fx> 100.0%
(0) <advisors, kept> -> <money-fx> 100.0%
(0) <ct, louisiana> -> <money-fx> 100.0%
(0) <algeria, buyout> -> <ship> 100.0%
(0) <bonds, delay> -> <ship> 100.0%
(0) <declaration, diagnostic> -> <ship> 100.0%
(0) <colombia, discount> -> <ship> 100.0%
(0) <bonds, employers> -> <ship> 100.0%
(0) <bonds, exported> -> <ship> 100.0%
(0) <diagnostic, feb> -> <ship> 100.0%
(0) <damaged, gone> -> <ship> 100.0%
(0) <agricultural, asks> -> <trade> 100.0%
(0) <adopt> -> <earn> 99.63%
(0) <chartered> -> <earn> 97.82%
    
```

```

Command Prompt - java TextMiningGUI_App
(0) <central> -> <ship> 66.66%
(0) <congressmen> -> <money-fx> 66.66%
(0) <agricultural, amc> -> <interest> 66.66%
(0) <exxon> -> <ship> 66.66%
(0) <exported> -> <trade> 66.66%
(0) <hearing> -> <money-fx> 66.66%
(0) <hands> -> <trade> 66.66%
(0) <hundreds> -> <trade> 66.66%
(0) <advisors, congressional> -> <interest> 66.66%
(0) <leveraged> -> <money-fx> 66.66%
(0) <live> -> <ship> 66.66%
(0) <livestock> -> <trade> 66.66%
(0) <attend, buys> -> <interest> 66.66%
(0) <auctions, buys> -> <interest> 66.66%
(0) <agriculture, citibank> -> <interest> 66.66%
(0) <announcing, commodity> -> <interest> 66.66%
(0) <attend, crops> -> <interest> 66.66%
(0) <commodity, dead> -> <interest> 66.66%
(0) <agriculture, effective> -> <money-fx> 66.66%
(0) <bonds, exchequer> -> <ship> 66.66%
(0) <principles> -> <acq> 66.66%
(0) <propose> -> <crude> 66.66%
(0) <prospect> -> <crude> 66.66%
(0) <protect> -> <crude> 66.66%
(0) <primarily> -> <earn> 66.66%
(0) <recommendations> -> <earn> 66.66%
(0) <reforms> -> <interest> 66.66%
(0) <pumping> -> <money-fx> 66.66%
(0) <rand> -> <money-fx> 66.66%
(0) <rapidly> -> <money-fx> 66.66%
(0) <quarter> -> <ship> 66.66%
(0) <raises> -> <ship> 66.66%
(0) <prior> -> <trade> 66.66%
(0) <ctuly> -> <trade> 66.66%
(0) <adjusted, algeria, attack> -> <crude> 66.66%
(0) <adjusted, chinese> -> <acq> 66.66%
(0) <agriculture, comprising> -> <acq> 66.66%
(0) <achieve, crude> -> <acq> 66.66%
(0) <ambassador, api> -> <crude> 66.66%
(0) <across, brazil> -> <crude> 66.66%
(0) <agricultural, brazilian> -> <earn> 66.66%
(0) <alvite, chipmakers> -> <earn> 66.66%
(0) <banks, chips> -> <grain> 66.66%
(0) <base, deferred> -> <grain> 66.66%
(0) <announcing, assumed> -> <interest> 66.66%
(0) <affected, bond> -> <interest> 66.66%
(0) <bodies, bureau> -> <interest> 66.66%
(0) <agriculture, closely> -> <interest> 66.66%
(0) <announcing, congressional> -> <interest> 66.66%
(0) <carryforwards, congressional> -> <interest> 66.66%
(0) <bond, congressmen> -> <interest> 66.66%
(0) <alvite, dan> -> <interest> 66.66%
(0) <associated, dan> -> <interest> 66.66%
(0) <congressmen, dan> -> <interest> 66.66%
(0) <bond, dead> -> <interest> 66.66%
(0) <consecutive, discovered> -> <interest> 66.66%
(0) <auctions, dispute> -> <interest> 66.66%
(0) <bond, dispute> -> <interest> 66.66%
(0) <dan, dispute> -> <interest> 66.66%
(0) <agriculture, easier> -> <interest> 66.66%
(0) <commodity, effective> -> <interest> 66.66%
(0) <attend, ends> -> <interest> 66.66%
(0) <advisors, followed> -> <interest> 66.66%
(0) <advisors, indicates> -> <interest> 66.66%
(0) <acres, assistance> -> <money-fx> 66.66%
(0) <assistance, bond> -> <money-fx> 66.66%
(0) <acquires, cheap> -> <money-fx> 66.66%
(0) <advisors, dealer> -> <money-fx> 66.66%
    
```

Some Interesting Rules

```

Command Prompt - java TextMiningGUI_App
*****
*
*          START TEST RUN
*
*****
SETTINGS
-----
Confidence (default 80%) = 35.0
Significance index       = 1.0E-4
Lower noise t'hold      = 0.2
Support (default 20%)   = 0.1
Upper noise t'hold      = 20.0
Maximum Num. Sig. Words = 1200

PHRASE MINING
Delimiters = stop marks and noise words
Contents   = at least one significant word and ordinary words

START ANALYSING DOC BASE FOR PHRASES

READ TRAINING SET
=====
Num. docs. in training set = 3321

LD NOISE, ORDINARY AND SIGNIFICANT WORDS IN WORDS BIM TREE
LNT = 0.2% (6 docs), UNT = 20.0% (664 docs), SI = 1.0E-4
Max # sig words 1200
-----
Significance contribution calculation strategy = Mutual Information
Potential Sig. Word. list generation strategy = All Significant Words
Sig. Word. identification strategy = First N (distributed), N is 1200 words.
Number of one itensets = 18649
READ TEST SET
=====
Num. docs. in test set = 3322

START CLASSIFICATION (BEST FIRST), TFPC WITH X-CHECKING AND CSA ORDERING
-----
Max num frequent sets = 500000
Max size of antecedent = 6
Number of records in training set = 3321
NOTE: Data set reordered
Creating P-tree table
Support = 0.1, Confidence = 35.0
Minimum support = 3.32 (Records)
Max num frequent sets = 500000
Max size of antecedent = 6
Number of records in training set = 3321
NOTE: Data set reordered
Apriori-TFP with X-Checking
Minimum support threshold = 0.1% (3.32 (records)
Generation time = 169.33 seconds (2.82 mins)
Accuracy = 07.21
Number of rules = 1083
Max size of antecedent = 2
Ave size of antecedent = 1.02
Antecedent size distribution
Antecedent size 1 = 1063
Antecedent size 2 = 20

CLASSIFIER
(0) <<above normal>> -> <earn> 100.0%
(0) <<accept ec>> -> <earn> 100.0%
(0) <<accept santos bid>> -> <earn> 100.0%
(0) <<accounted>> -> <earn> 100.0%
(0) <<acquire another>> -> <earn> 100.0%
(0) <<acquire any additional staff>> -> <crude> 100.0%
(0) <<acquire chemlawn>> -> <acq> 100.0%
(0) <<acquire coastal bancorp>> -> <crude> 100.0%
(0) <<acquire existing investment dealers>> -> <earn> 100.0%
(0) <<acquire john paul>> -> <earn> 100.0%
(0) <<acquire new stock representing>> -> <earn> 100.0%
(0) <<acquire north>> -> <earn> 100.0%
(0) <<acquire rest>> -> <earn> 100.0%
(0) <<acquire some>> -> <earn> 100.0%
(0) <<acquire stock>> -> <earn> 100.0%
(0) <<acquired business>> -> <acq> 100.0%
(0) <<acquire united national bank>> -> <earn> 100.0%
(0) <<acquired all issued shares>> -> <earn> 100.0%
(0) <<acquired electronics components>> -> <acq> 100.0%

```

LUCS-KDD: Text Mining GUI

File Thresholds Sig. Word Strats. Algorithm Evaluation Class. Strat. Output Batch

About

Text Base Dir. C:\3

Text Base Spec. ALUES:

First N recs ---

Start ALUES:

Start Batch Mode

Start Set of 8 Mode

Exit ---

SET MAXIMUM NUMBER SIGNIFICANT WORDS:
Maximum number of permitted significant words = 1200

Selected Significance contribution calculation strategy = Mutual Information

Selected potential significance words list generation strategy = All Significant Words

Selected significant word ID strategy = First N (distributed)

Selected TM algorithm = Keywords

Selected rule evaluation strategy = 50:50

Output:
Classifier Stats. = true

Output:
Classifier Stats. = true
Classifier = true

START CLASSIFICATION:
Text Base Dir. = C:\Documents and Settings\Yanbo J. Wang\Java Code\PHD_related_software\TPP\Input_Re
Class list = {acq, crude, earn, grain, interest, money-fx, ship, trade}
FileStem = Reuters
FileEnd = .txt
Num. records = 6643
Support = 0.1
Confidence = 35.0
Significance = 1.0E-4
Upper Noise = 20.0
Lower Noise = 0.2
Sig. Wd. strat. = First N (distributed)
TM algorithm = Keywords
Sat. Strategy = CSA
Eval. strat. = 50:50

Selected TM algorithm = DelSN_contGO

LUCS-KDD (Liverpool University Computer Science - Knowledge Discovery
in Data) group Text Mining demonstrator.

Version 1 Created by Frans Coenen (28 February 2006)

The Phrase Approach


```

Command Prompt - java TextMiningGUI_App
(0) <<agricultural options>> -> {crude} 100.0%
(0) <<agree oil exploration project>> -> {earn} 100.0%
(0) <<agreed production ceiling>> -> {earn} 100.0%
(0) <<against new york developer donald trump>> -> {money-fx} 100.0%
(0) <<against oil>> -> {money-fx} 100.0%
(0) <<agricultural futures markets>> -> {money-fx} 100.0%
(0) <<agricultural legislation>> -> {ship} 100.0%
(0) <<agricultural policy>> -> {acq} 100.0%
(0) <<agricultural products businesses>> -> {acq} 100.0%
(0) <<agriculture if>> -> {acq} 100.0%
(0) <<agriculture secretary richard>> -> {acq} 100.0%
(0) <<aid program outside>> -> {acq} 100.0%
(0) <<air atlanta could sell>> -> {acq} 100.0%
(0) <<air canada courier buy sharply>> -> {acq} 100.0%
(0) <<air north america provided no support>> -> {acq} 100.0%
(0) <<agriculture department currently forecasts this>> -> {crude} 100.0%
(0) <<agriculture ministry official told reuters>> -> {crude} 100.0%
(0) <<agricultural production>> -> {earn} 100.0%
(0) <<agricultural trade reform under>> -> {earn} 100.0%
(0) <<agriculture department analysts>> -> {earn} 100.0%
(0) <<agriculture department officials>> -> {earn} 100.0%
(0) <<air quality>> -> {earn} 100.0%
(0) <<agricultural products>> -> {interest} 100.0%
(0) <<agricultural production sharply over>> -> {money-fx} 100.0%
(0) <<agriculture subsidies>> -> {ship} 100.0%
(0) <<agriculture undersecretary daniel>> -> {ship} 100.0%
(0) <<aircraft engine repair>> -> {acq} 100.0%
(0) <<airline reported provisional>> -> {acq} 100.0%
(0) <<alaska should>> -> {acq} 100.0%
(0) <<all crude>> -> {acq} 100.0%
(0) <<all gencorp>> -> {acq} 100.0%
(0) <<all united airlines flights>> -> {acq} 100.0%
(0) <<allegheeny board>> -> {acq} 100.0%
(0) <<allegheeny made>> -> {acq} 100.0%
(0) <<allegheeny voting power>> -> {acq} 100.0%
(0) <<allied irish banks foreign exchange dealer john>> -> {acq} 100.0%
(0) <<allow consumers there>> -> {acq} 100.0%
(0) <<alaska>> -> {crude} 100.0%
(0) <<all shareholders other than norfolk southern>> -> {crude} 100.0%
(0) <<alleged european government subsidies>> -> {crude} 100.0%
(0) <<allegheeny>> -> {crude} 100.0%
(0) <<alaska north>> -> {earn} 100.0%
(0) <<all farm subsidies affecting trade within ten years>> -> {earn} 100.0%
(0) <<all federal waiting period requirements>> -> {earn} 100.0%
(0) <<all other gatt member>> -> {earn} 100.0%
(0) <<all shares tendered>> -> {earn} 100.0%
(0) <<all transactions>> -> {earn} 100.0%
(0) <<allegheeny reported earnings>> -> {earn} 100.0%
(0) <<allegheeny spokesman>> -> {earn} 100.0%
(0) <<allies look very strongly>> -> {earn} 100.0%
(0) <<allegheeny county>> -> {interest} 100.0%
(0) <<algeria have already called>> -> {money-fx} 100.0%
(0) <<all funds under management>> -> {money-fx} 100.0%
(0) <<all grades>> -> {money-fx} 100.0%
(0) <<all opec countries>> -> {money-fx} 100.0%
(0) <<all tender offers should remain open>> -> {money-fx} 100.0%
(0) <<all these bilateral>> -> {money-fx} 100.0%
(0) <<allow canadian banks>> -> {money-fx} 100.0%
(0) <<allow gcc citizens>> -> {money-fx} 100.0%
(0) <<allow exports>> -> {trade} 100.0%
(0) <<allow monetary policy>> -> {trade} 100.0%
(0) <<allowed imports>> -> {acq} 100.0%
(0) <<already caused losses estimated>> -> {acq} 100.0%
(0) <<already lowered>> -> {acq} 100.0%
(0) <<already owns>> -> {acq} 100.0%
(0) <<ambassador real estate>> -> {acq} 100.0%
(0) <<anc common shares opened>> -> {acq} 100.0%
(0) <<anc shareholders other than renault>> -> {acq} 100.0%

```

```

Command Prompt - java TextMiningGUI_App
(0) <<allies must>> -> {crude} 50.0%
(0) <<algeria iranian oil minister>> -> {earn} 50.0%
(0) <<all cases unchanged>> -> {earn} 50.0%
(0) <<all outstanding cyclops shares>> -> {interest} 50.0%
(0) <<all properties>> -> {interest} 50.0%
(0) <<allegheeny international>> -> {interest} 50.0%
(0) <<all outstanding cyclops shares>> -> {money-fx} 50.0%
(0) <<allies must>> -> {money-fx} 50.0%
(0) <<allow all exploration expenditure>> -> {money-fx} 50.0%
(0) <<allegheeny noted>> -> {trade} 50.0%
(0) <<allow ecuador>> -> {trade} 50.0%
(0) <<acquire taft broadcasting despite>> -> {acq} 48.27%
(0) <<acquired utah>> -> {interest} 47.82%
(0) <<accord last month reached>> -> {money-fx} 47.16%
(0) <<acquisition transactions>> -> {money-fx} 47.05%
(0) <<acquisition requires>> -> {interest} 47.05%
(0) <<acquisition until negotiations are completed>> -> {earn} 47.05%
(0) <<adding these cuts are expected>> -> {money-fx} 46.66%
(0) <<actual sales might exceed output due>> -> {acq} 46.66%
(0) <<advertisement placed>> -> {acq} 46.15%
(0) <<accused>> -> {earn} 45.67%
(0) <<acquire restaurants national>> -> {earn} 45.45%
(0) <<against earnings no later than january>> -> {acq} 45.45%
(0) <<after financial income>> -> {earn} 45.45%
(0) <<after real tax including>> -> {interest} 45.45%
(0) <<against international market turbulence>> -> {money-fx} 45.45%
(0) <<acquisition>> -> {interest} 45.0%
(0) <<acquire first>> -> {money-fx} 44.44%
(0) <<acquisition if subject>> -> {ship} 44.44%
(0) <<acquisition if subject>> -> {crude} 44.44%
(0) <<aide>> -> {trade} 44.44%
(0) <<agriculture negotiations reach>> -> {trade} 44.44%
(0) <<agriculture ministry sources>> -> {trade} 44.44%
(0) <<agriculture ministry officials>> -> {trade} 44.44%
(0) <<agriculture department>> -> {money-fx} 44.44%
(0) <<agricultural supports>> -> {money-fx} 44.44%
(0) <<aide indicated>> -> {grain} 44.44%
(0) <<aid program>> -> {earn} 44.44%
(0) <<air problems facing commerce>> -> {crude} 44.44%
(0) <<aide>> -> {acq} 44.44%
(0) <<agriculture department should make grain>> -> {acq} 44.44%
(0) <<agricultural producers>> -> {acq} 44.44%
(0) <<acquire hughes>> -> {crude} 43.9%
(0) <<acquire southern pacific>> -> {earn} 43.75%
(0) <<act between defending higher oil prices>> -> {money-fx} 43.75%
(0) <<adequate certificate>> -> {interest} 42.85%
(0) <<adjusted>> -> {earn} 42.85%
(0) <<adjust ecuador oil facility>> -> {earn} 42.85%
(0) <<acquire ecuador>> -> {crude} 41.66%
(0) <<acquire taft broadcasting co., {active institutional dollar sales}>> -> {money-fx} 41.66%
(0) <<after excluding coffee>> -> {trade} 41.66%
(0) <<affect both pretax>> -> {trade} 41.66%
(0) <<affiliated operations>> -> {ship} 41.66%
(0) <<after capital expenditure>> -> {money-fx} 41.66%
(0) <<affect drilling done>> -> {money-fx} 41.66%
(0) <<after capital expenditure>> -> {acq} 41.66%
(0) <<acquire supermarkets general>> -> {money-fx} 41.37%
(0) <<acquisition should have>> -> {money-fx} 41.17%
(0) <<acquisition raises>> -> {interest} 41.17%
(0) <<acquisition vote>> -> {acq} 41.17%
(0) <<acquire major>> -> {acq} 41.02%
(0) <<accounting method>> -> {money-fx} 40.62%
(0) <<added no extraordinary opec conference>> -> {money-fx} 40.0%
(0) <<additional crude oil storage>> -> {money-fx} 40.0%
(0) <<additional crude oil loan>> -> {ship} 40.0%
(0) <<accept credit guarantees, {acquire four florida banks}>> -> {acq} 40.0%

```

Some Interesting Rules

Stage-two Usage of TC in Intute

Stage-two Usage: *Multi-label TC*

Usually, a search result is presented as a (long) list of “matching” documents. **Fig. 2** shows the result for querying “*fuel crisis*” on Google. There are total 1,320,000 records returned. Obviously, no one will read them all. Hence presenting this search result in groups, separated by different topics (sub-domain-classes) is suggested.

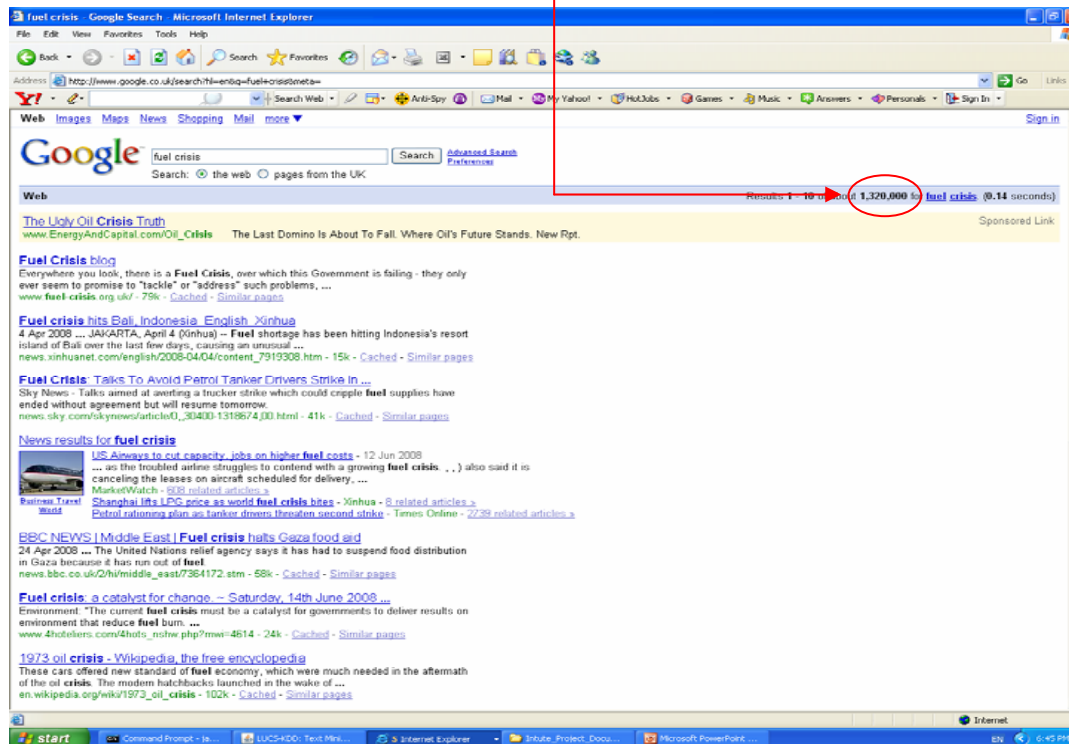


Fig. 2. A Search Result from Google

Stage-two Usage of TC in Intute

Broadly speaking, *Social Science* sub-branches include Anthropology, Economics, Education, Geography, History, Law, Linguistics, Political Science, Psychology, Social Work, Sociology, etc. Hence the search result of “*fuel crisis*” can be presented regarding these branch-classes (see **Fig. 3**). Note that a result document (record) may be associated with more than one branch-classes.

<u>Economics</u>	<u>Political Science</u>	<u>Geography</u>	<u>Law</u>
Document # 1	Document # 2	Document # 1	Document # 5
Document # 3	Document # 5	Document # 6	Document # 21
Document # 5	Document # 8	Document # 21	...
Document # 10	Document # 14	...	
...	...		

Fig. 3. Presenting a Search Result in Classes

Strategy of Multi-label TC

From the demo of Single-label TC, we see two rules as follows.

```
(0) {amc, buys} -> {interest} 80.0%
(0) {advisors, comprising} -> {interest} 80.0%
(0) {announcing, crops} -> {interest} 80.0%
(0) {citibank, crops} -> {interest} 80.0%
(0) {advisors, dec} -> {interest} 80.0%
(0) {agricultural, decade} -> {interest} 80.0%
(0) {bond, decade} -> {interest} 80.0%
(0) {dead, effective} -> {interest} 80.0%
(0) {bond, fairly} -> {interest} 80.0%
(0) {amc, gasoline} -> {interest} 80.0%
(0) {awabia, cheap} -> {money-fx} 80.0%
(0) {advisors, completes} -> {money-fx} 80.0%
(0) {advisors, consistent} -> {money-fx} 80.0%
(0) {closes, dependence} -> {money-fx} 80.0%
(0) {advisors, fire} -> {money-fx} 80.0%
(0) {announcing, follow} -> {money-fx} 80.0%
(0) {ct, halt} -> {money-fx} 80.0%
```

```
(0) {amc, crops} -> {interest} 100.0%
(0) {auctions, crops} -> {interest} 100.0%
(0) {citibank, effective} -> {interest} 100.0%
(0) {advisors, finalized} -> {interest} 100.0%
(0) {bodies, indicates} -> {interest} 100.0%
(0) {bureau, indicates} -> {interest} 100.0%
(0) {agriculture, arab} -> {money-fx} 100.0%
(0) {accord, cereals} -> {money-fx} 100.0%
(0) {advisors, classes} -> {money-fx} 100.0%
(0) {advisors, closes} -> {money-fx} 100.0%
(0) {agriculture, closes} -> {money-fx} 100.0%
(0) {advisors, completing} -> {money-fx} 100.0%
(0) {agriculture, containers} -> {money-fx} 100.0%
(0) {agriculture, dependence} -> {money-fx} 100.0%
(0) {advisors, discussion} -> {money-fx} 100.0%
(0) {advisors, kept} -> {money-fx} 100.0%
(0) {ct, louisiana} -> {money-fx} 100.0%
```

Hence we indicate that a compound rule can be described as:

{Advisors, Completes/Completing} ⇒ {money-fx}

Strategy of Multi-label TC

Also from the demo of Single-label TC, we see another two rules.

```

(0) {implement} -> {trade} 71.42%
(0) {initially} -> {money-fx} 71.42%
(0) {industrialised} -> {money-fx} 71.42%
(0) {improving} -> {earn} 71.42%
(0) {bilateral, boat} -> {trade} 71.42%
(0) {agriculture, fire} -> {money-fx} 71.42%
(0) {auditors, discontinued} -> {money-fx} 71.42%
(0) {advisors, bonds} -> {money-fx} 71.42%
(0) {advisors, fluctuate} -> {interest} 71.42%
(0) {announcing, fairly} -> {interest} 71.42%
(0) {amc, fairly} -> {interest} 71.42%
(0) {announcing, dan} -> {interest} 71.42%
(0) {amc, commodity} -> {interest} 71.42%
(0) {auditors, bond} -> {interest} 71.42%
(0) {agriculture, alvite} -> {interest} 71.42%
(0) {arab, deferred} -> {grain} 71.42%
(0) {zambia} -> {trade} 71.42%

```

```

(0) {fruit} -> {crude} 68.75%
(0) {forward} -> {money-fx} 68.75%
(0) {foundation} -> {money-fx} 68.75%
(0) {gatt} -> {trade} 68.75%
(0) {gnp} -> {trade} 68.75%
(0) {advisors, citibank} -> {interest} 68.42%
(0) {engine} -> {crude} 68.42%
(0) {advisors, bond} -> {interest} 68.0%
(0) {allegheny} -> {earn} 67.74%
(0) {cope} -> {trade} 67.74%
(0) {achieve} -> {earn} 67.62%
(0) {asks} -> {earn} 67.56%
(0) {api} -> {earn} 67.17%
(0) {central} -> {ship} 66.66%
(0) {congressmen} -> {money-fx} 66.66%
(0) {agricultural, amc} -> {interest} 66.66%
(0) {exxon} -> {ship} 66.66%

```

Hence we indicate that a multi-labeled compound rule can be described as:

{Advisors, Bonds/Bond} ⇒ {money-fx, interest}

Further Development

Fig. 4 shows the HASSET (Humanities and Social Science Electronic Thesaurus) categories. The HASSET categories can be used to present *Social Science* related documents in subject/domain hierarchies. We introduce an hierarchical multi-label TC problem to map new unlabeled documents to the HASSET hierarchy. This allows the user to concentrate on a “small” group of “interesting” results and offers a solution to the problem of information overload.

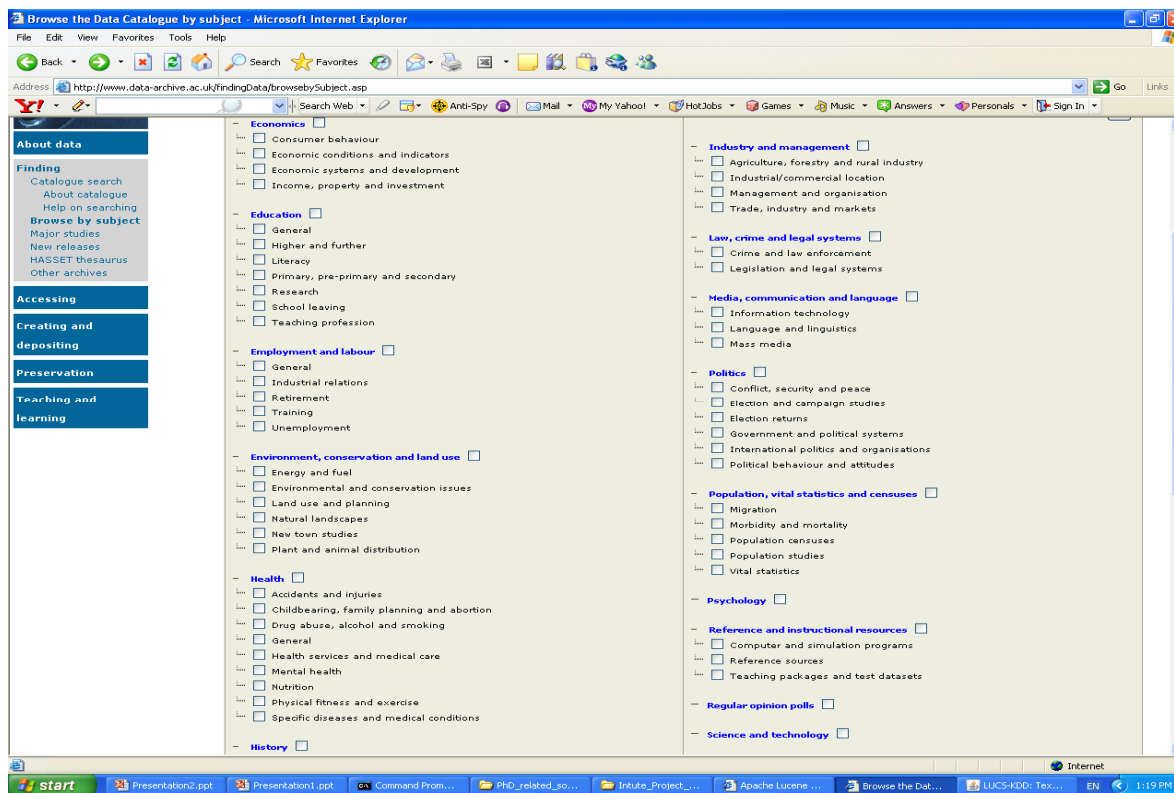


Fig. 4. The HASSET Categories

Summary

The Intute project aims to develop an intelligent semantic web search system that deals with *Social Science* and Bio-medical Science documents.

Text classification is a well-known research area that maps documents to pre-defined categories. More than this, the techniques we use allow users to see why those predictions have been made.

As work continues on the Intute project, we will be adding a number of other text mining tools to support cross-repository search focusing on areas of interest to social scientists.

Questions?