Offline Evaluation of Term Utility Functions

Alexander K. Seewald, **Christian Holzbaur** and **Gerhard Widmer**
Austrian Research Institute for Artificial Intelligence
Schottengasse 3, A-1010 Wien, Austria
[alexsee,christian,gerhard]@ai.univie.ac.at

Abstract

In this paper we investigate characterizing ontology nodes corresponding to human-comprehensible concepts from the tool Melvil by a set of terms. We choose a variety of term utility functions, commonly use in text mining, to determine relative importance of terms for the task of deciding if a given document is part of a certain concept or not. We evaluated each utility function both quantitatively by considering precision and recall of the top ten terms returned and qualitatively by analyzing which of the original patterns and obviously related terms were recovered. This approach could be used to suggest promising terms to a human ontology editor during creation of a new node. Our results look somewhat promising but still needful of improvement – so we also report on probable causes of unsatisfactory results.

1 Introduction

The tool melvil allows to create ontology nodes, each one based on a human-comprehensible concept, and organize them in an arbitrary hierarchy. Each ontology node has an associated regular expression pattern which is used to retrieve all relevant documents of the corresponding concept. A concept such as *Internet* may have a pattern such as `\binternet`[1]`\b|\bweb\b|\bwww\b`, which consists of several subpatterns separated by `|`. Not all of these patterns must be single words as in our example – multiword patterns also appear frequently, at least in the research ontology which was provided by uma. Notice also that `\b` is used to denote word boundaries – in our example, the subpatterns match whole words only.

In this paper, we investigate characterizing concepts by a short list of appropriate terms, i.e. single words. We consider a variety of term utility functions, each of which map every tuple (term, concept) to a numeric value which signifies the usefulness of the respective term to decide if documents are part of the respective concept or not.

Such a characterization by terms could also suggest new single word patterns to the user during creation of a new ontology node, if real-time performance is achievable.

We will first give an overview about the research ontology, March15, which was provided by uma, focussing on common data statistics. Afterwards, we will introduce the term utility functions used throughout our experiments. These are more concisely referred to as *measures*. Then we will shortly describe the experimental setup, discuss major experimental results in the Results section and discuss minor experimental results and other issues in the Discussion section. At last, we will conclude this paper.

[1] i.e. the term *internet*.

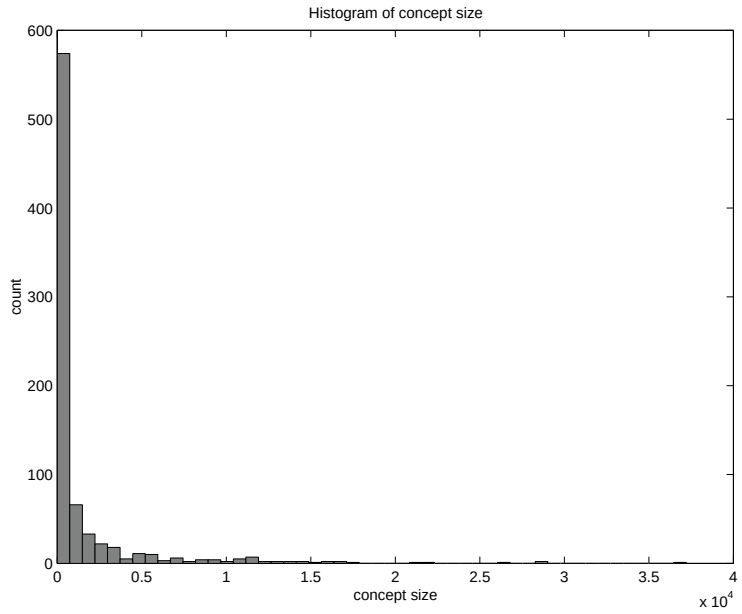


Figure 1: This is a histogram of concept size. Notice that the concept size is scaled by 10^4, so a value of 1 stands for 10,000 documents on the x-axis.

2 Data Analysis

The March15 ontology contains 793 nodes resp. concepts. On average, these concepts contain 1,395.5±3,583.9 documents – Figure 2 shows a histogram of concept size. The top ten concepts account for 230,161 documents[2]. Since there are only 55,047 unique documents, this means that most documents are assigned to many concepts. The overlap[3] for these ten concepts alone is already 4.18. About three quarters (76.0%) of all concepts contain less than a thousand documents; there are even 61 concepts which do not contain any documents! This exponential falloff in concept size is responsible for the mentioned abnormally high standard deviation.

Hierarchical relations between concepts can be modelled by the user – however, local regular expression at every node are used to determine the documents which are part of each node, independent of all other nodes; so this hierarchy is not relevant for the purposes of our report.[4] Nevertheless, we have noted that the given hierarchy is not a forest[5] as usually expected from an ontology but a acyclic[6] graph, e.g. node *sms* has four parent nodes. Many nodes are thus reachable on multiple paths. Some top-level nodes such as *altavista*, *region* and *technology* do not contain any documents and seem to have been created simply as container for children nodes.

There are 55,047 unique documents altogether which usually belong to multiple concepts – the average overlap over all concepts is 20.08. The documents are partitioned into 187

[2] 20.8% of the combined size of all concepts, which is 1,105,270.

[3] $overlap = \frac{\sum conceptSize(c)}{number\ of\ unique\ documents}$

[4] The term *ontology* may thus be considered slightly inappropriate in the current implementation of melvil.

[5] Forest = a set of disconnected trees. A tree is a acyclic graph where every node has at most one parent node.

[6] Cycles seem possible in practice, if the ontology editor does not prevent this.

	t	¬ t
c	a	b
¬c	c	d

Table 1: This is the contingency table for term *t* and concept *c*. a,b,c,d are the number of documents in the four categories along two independent dimensions: term occurrence and concept membership.

jobs[7] where every job has its own fulltext index so each has to be accessed separately. Three quarters (74.8%) of all jobs contain less than 100 documents. Due to resource constraints, we were forced to restrict our experiments to the largest ten jobs, which account for 69% of all unique documents.

A total of 428,173 terms were indexed. This large number can be explained by the existence of multiple languages within our corpus and the use of all alphanumeric patterns for indexing. When using only the largest ten jobs, 32.6% of all terms never appear; 86.5% appear in less than ten documents and 96.6% appear in less than a hundred documents[8]. When looking closely at these terms, we have found that quite many of them may be explained as internal html-tags[9] or other suspicious words[10] which are usually not considered to be part of human-readable text.

We have also found that, contrary to our previous assumption, the full text index seems to be based on substring search so that e.g. *in* is found both as a single word and as part of larger words such as *internet*[11]. On the other hand, not all substrings were indexed, e.g. *rnet* does not appear as a term. We therefore presume that the terms were initially constructed by parsing the documents while considering word boundaries – however the full text index was later generated by searching for these terms as substrings. Thus, when searching for term *web*, other terms such as *schwebend*, *textilgewebe* and *feldwebel* may also contribute matching documents which would be inappropriate.

3 Term Utility Functions

In our experiments, we used the following measures of term utility, most of which are commonly used for text-mining. All but two of these measures can be computed from the contingency table which is described in Table 3. Additionally, since the contingency table only captures term occurrence, we also calculated the average term frequency for documents[12] inside (sfC) and outside ($\neg sfC$) the concept. These were used for measures *sR2* and *sR3*. The following measures were tested:

- χ^2 which determines if there is a statistically significant relation between term occurrence and concept membership (Yang & Pedersen, 1997).
- Information Gain (*IG*) which determines the information gained for concept prediction, given term occurrence. Both IG and χ^2 were found to be superior to all other considered features in (Yang & Pedersen, 1997).

[7] Every job corresponds to an internet or intranet server where all its associated documents are fetched from.

[8] 99.4% appear in less than a thousand documents.

[9] e.g. td, 7pt, mediumbold, boldlink, ft26xx3044x11, writelayersn...

[10] e.g. 00000000000000000000000001, aaaaaa, abcdefghijklmnopqrstuvwxyz...

[11] If it were different, combining the *internet* node contingency tables of terms *www*, *web* and *internet* would yield a combined table with precision and recall both 1.0. In fact, p=0.91 and r=0.52!

[12] We considered only of those documents where the term appears at least once; i.e. cases a & c in the contingency table. The average term frequency for the other two cases is obviously zero.

conceptId	*numDocs*	*numPatts*	*numKids*	*numParents*	*level(s)*
internet	37211	3	6	2	1,4
e_mail	28372	2	0	2	4,7
business unit	28360	4	2	1	1
search	26270	2	5	2	3,4,5,6,8
mobile	22035	4	9	1	2
media	21551	3	8	1	4,6,7
software	17119	1	6	1	2
mobile phone	16654	4	0	1	4
investment/investor	16462	2	2	1	4
network	16127	2	5	1	3,6
artifi cial intelligence	239	4	5	2	3,6
online community	350	3	0	1	5
manufacturer	379	4	4	1	3
wlan	974	3	0	1	4
market capitalisation	504	2	0	1	4
cryptography	831	3	0	1	4
mobile portal	474	4	13	2	3,5,6
online gaming	534	2	0	3	4,5,6,7
knowledge management	758	3	0	2	4,5,6,7,9
price earning ratio	392	4	0	1	4

Table 2: This table shows the ontology nodes resp. concepts chosen for our experiments. The top half shows the large concepts, the bottom half the small ones. *numPatts* is the number of patterns used for indexing. *numKids* and *numParents* shows the number of children resp. parent nodes. Level(s) shows the zero-based level index for each concept – if a node is accessible via multiple paths and/or multiple root nodes, more than one level has to be shown.

- *oddsRatio*, which is a commonly used feature in information retrieval (van Rijsbergen, Harper & Porter, 1981). In our case, when removing the logarithm which is irrelevant for relative ranking of terms, this simplifies to $\frac{ad}{bc}$.
- *odds2* is one of the many measures inspired by the original Odds Ratio formula, i.e. $\frac{a+c}{N}\log_2\frac{ac+ad}{ac+bc}$ where $N = a + b + c + d$ is the total number of documents. It is equivalent to *FreqLogP* in (Mladenic, 1998). *odds2* and *oddsRatio* were two of three measures found to be superior in the mentioned paper. The third measure was based on exponentiation and deemed to be too costly for implementation.
- Precision (*prec*) – the ratio of documents belonging to the concept among all documents which include the term, i.e. $\frac{a}{a+c}$.
- Recall (*recall*) – the ratio of documents which include the term, among all documents belonging to the concept, i.e. $\frac{a}{a+b}$.
- $prec * recall$ (*PR*) which trades off recall and precision. Usually, we want both high precision and high recall – our formula is a simple way to capture this relation.
- SimpleRatio1 (*sR1*) is $\frac{a}{N}$ where $N = a+b+c+d$ is the overall number of documents.
- SimpleRatio2 (*sR2*) is $\frac{sfC}{N}$, which prefers those terms which appear very frequently.
- SimpleRatio3 (*sR3*) is $\frac{sfC}{\neg sfC+1}$, which prefers those terms appearing frequently within the concept, but seldom without.

In initial experiments we found that *prec* is unusable since about 20% of the terms we looked at have the maximum precision of $prec = 1.0$ ($a > 0$ and $c = 0$), which makes it impossible to determine a stable relative ranking![13] We also found that *sR1* and *sr2* perform very poorly, mostly selecting stop words. So we removed the three mentioned measures which still leaves us with seven measures for offline our evaluation. A simplistic way to evaluate these term utility measures is to look at which terms correspond to the indexing patterns. But since using single terms instead of regular expressions is a crude approximation at best, some precision is inevitably lost. So we used precision and recall, averaged over the top ten terms, as fairer evaluation.

4 Experimental Setup

As we already mentioned, due to resource constraints we used only the largest ten jobs[14] in our experiments. These account for 69% of all unique documents. For the same reason we were unable to evaluate all 793 concepts, so we chose twenty concepts for further investigation: the ten largest concepts by size and also ten smaller concepts with 200-1000 indexed documents which were arbitrarily chosen. Details can be found in Table 3.

For these twenty concepts, we computed the contingency tables plus sfC and $\neg sfC$ for each term, by summing over all chosen jobs.[15] This data was extracted from Melvil and then imported into MATLAB for further offline analysis, including computing all our term utility functions, statistical analysis and visualization. In MATLAB, computing all seven measures for ten concepts and all 428,173 terms took about thirty seconds.

5 Results

We first determined the top ten terms for each measure – these are shown in Tables 8 and 8 in the Appendix. Terms which are matched by the original indexing patterns are shown in **bold**. Multiword patterns were broken up into single word patterns at every place where a word boundary may appear, e.g. `e.?mail` maps to `e|mail|email`. For all but one concepts, at least one indexing pattern is recovered. Detailed qualitative results are discussed in the next section.

We then computed precision and recall for each concept and measure, averaged over the top ten terms. The complete results can be found in Table 5. For comparing the different measures at one glance, we have also computed the grand mean and standard deviation of recall and precision over large and small concepts separately, these are visualized in Figure 4. High precision and high recall – which is what we usually want – is found towards the top right of each graph. Detailed plots which show the top ten terms for each concept and measure can be found in the Appendix.

All measures seem to work quite well on our ten large concepts, but perform more poorly on the ten small ones. However, the relative order between measures is still quite similar – two adjacent measures (*IG*=$\circ$, *PR*=$\square$) switch places and *odds2* ($*$) moves from top left (high precision, low recall) to bottom right (low precision, high recall).

Using *PR*, i.e. *prec* multiplied by *recall* seems to work quite well, so we have extended the formula to $prec^x * recall^{2-x}$ and plotted grand mean precision and recall for ten different

[13] I.e. we would have to randomly choose ten best terms out of about 80,000 equally good ones.

[14] jobIds = 0,14,100,91,97,16,13,94,92,2

[15] We also investigated normalizing the job size prior to combining the contingency tables – the results were mixed, see Discussion.

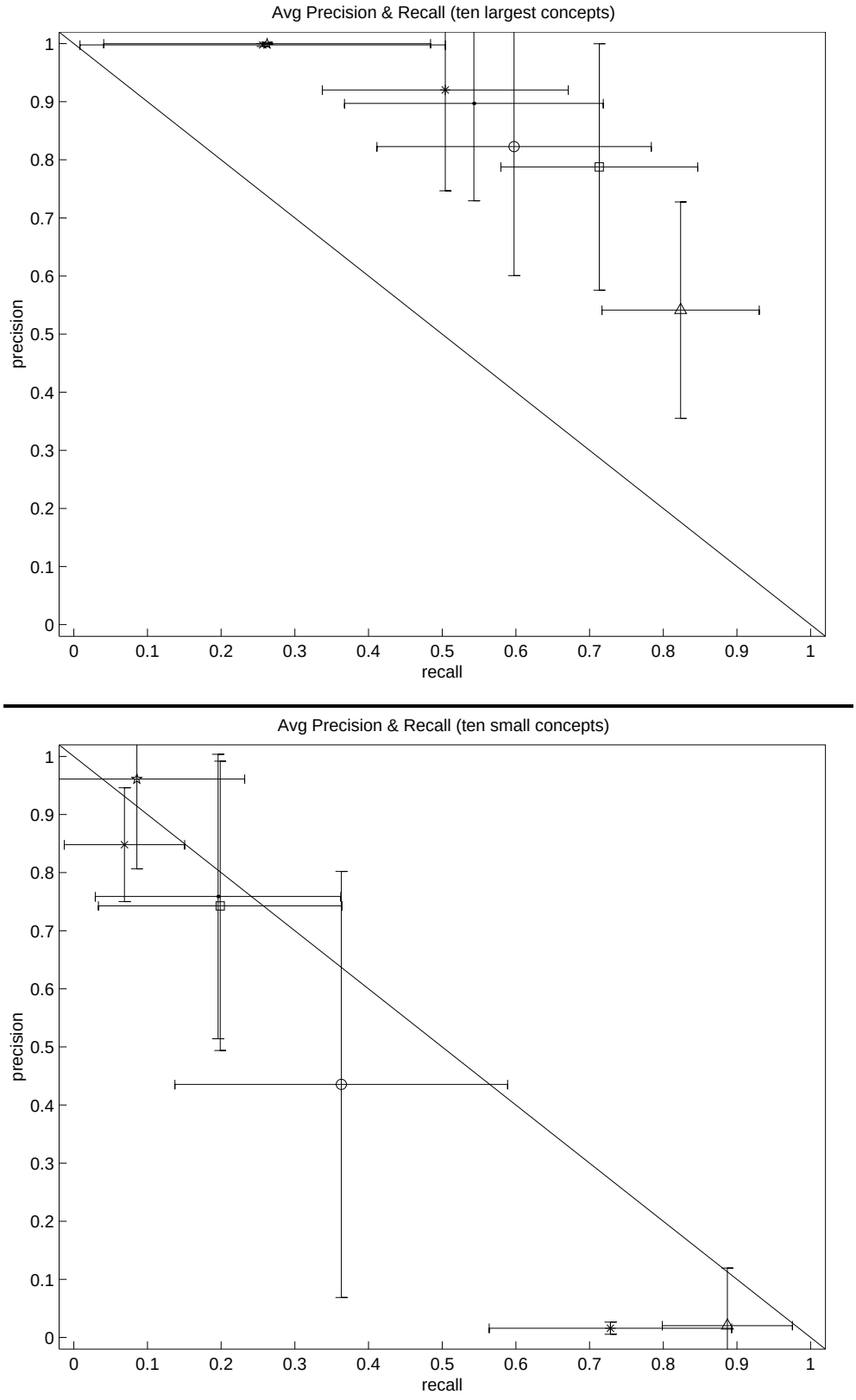


Figure 2: These plots show the average precision and recall $\pm$ standard deviation for the ten largest concepts (top) and our ten small concepts (bottom) – one point from each measure, averaged over the top ten terms. $\bullet = \chi^2$, $\circ$=*IG*, $\times$=*oddsRatio*, $*$=*odds2*, $\square$=*PR*, $\triangle$=*recall* and $\star$=*sR3*.

values of x in Figure 5. This gives competitive results and also allows for user-defined explicit trade-off between precision and recall. For $x = 0$ the formula is equivalent to *recall* and for $x = 1$ to *PR*. x should always be smaller than 2 because then the formula would be equivalent to *prec* and $prec = 1.0$ for about 20% of all terms as we mentioned earlier, so the obtained ranking would be quite random.

6 Discussion

In Table 8 concerned with the large concepts, some indexing patterns are recovered and usually appear near the top. Also, some obviously related patterns are found, e.g. *sms* and *wap* for concept handy, *kapital* for business unit and *microsoft* for software which is even found by more than half of our measures, always on second rank after *software* itself. For concept search, no indexing pattern could be recovered, but at least some related terms such as *recherche* and *stichwortsuche* are found.

In Table 8 concerned with the small concepts, we see a similar picture: some indexing patterns are recovered – sometimes even quite many e.g. for concepts manufacturer and cryptography. Related terms are also quite apparant: e.g. *blackbox* for concept community, *802 11a*[16] for concept WLAN, many names of popular online games for concept online gaming and *kampfroboter*, *wunderwaffe* and *arbeitssklaven* for concept artificial intelligence. We were also quite surprised to notice *trappl* as third-most relevant term for concept artificial intelligence by measure *sR3* which otherwise performs poorly. Another less pleasant surprise was that the term *fuck* appears on second place in three measures for concept online community. When we look at precision and recall, we see that other top terms such as *blackbox* (p=0.82, r=0.18), *community* (p=0.71, r=0.14) and *fuck* (p=0.72, r=0.19) seem quite similar. Although community was used as indexing pattern for this concept, its precision is less than 1.0 – another indication that the full text index does not index whole words as we mentioned earlier.

When we normalized the job size before combining the respective contingency tables, we encountered mixed results: On the one hand, the terms corresponding to indexing patterns sometimes move more to the top[17], and sometimes one or more additional indexing pattern are found[18] On the other hand, up to four indexing patterns[19] are lost and many of the related words disappear. It seems that this variation works better on the large concepts than on the small ones. Details can be found in Tables 8 and 8. The average precision and recall is very similar between normalized and non-normalized case – in fact, the grand average precision and recall plots were so similar to Figure 5 as to be almost indistinguishable. So we are inclined to prefer the original non-normalized version by Ockham's Razor.

For our experiments, computing the contingency table, sfC and $\neg sfC$ took an average of 6.75 minutes per concept. To achieve realtime performance, a computational speedup of at least two orders of magnitude is necessary. Notice also that we only used the top ten jobs in our experiments instead of all 187 which may be problematic when very heterogenous document sources are employed.

A possible way to achieve this would be to precompute these values efficiently. This would also results in an additional advantage would be that from our six values, many basic statistics such as precision, recall, overall term frequency and term frequency within con-

[16] IEEE 802.11a WLAN standard

[17] mobile, media, network, price earning ratio

[18] e_mail, investor, artifi cial intelligence, search

[19] for concept manufacturer and measure χ^2

Measure	*conceptId*	*avg prec.*	*avg recall*
χ^2	internet	0.947±0.041	0.457±0.160
	e_mail	0.765±0.296	0.550±0.237
	business unit	0.979±0.060	0.426±0.163
	search	0.900±0.079	0.642±0.129
	mobile	0.963±0.053	0.618±0.060
	media	0.958±0.060	0.581±0.123
	software	0.625±0.131	0.513±0.184
	mobile phone	0.985±0.021	0.650±0.091
	investment/inv.	0.855±0.209	0.374±0.238
	network	0.993±0.010	0.619±0.048
IG	internet	0.934±0.045	0.479±0.148
	e_mail	0.765±0.296	0.550±0.237
	business unit	0.687±0.215	0.719±0.261
	search	0.803±0.238	0.632±0.211
	mobile	0.929±0.074	0.644±0.050
	media	0.955±0.059	0.583±0.122
	software	0.602±0.142	0.551±0.171
	mobile phone	0.963±0.068	0.667±0.094
	investment/inv.	0.614±0.306	0.523±0.281
	network	0.975±0.035	0.627±0.048
oddsRatio	internet	0.999±0.000	0.126±0.100
	e_mail	0.998±0.000	0.074±0.075
	business unit	0.999±0.000	0.262±0.238
	search	0.999±0.000	0.267±0.134
	mobile	0.997±0.001	0.274±0.247
	media	0.999±0.001	0.329±0.201
	software	0.994±0.005	0.115±0.299
	mobile phone	0.999±0.002	0.568±0.189
	investment/inv.	0.995±0.003	0.121±0.200
	network	0.997±0.002	0.426±0.275
odds2	internet	0.995±0.005	0.330±0.170
	e_mail	0.976±0.026	0.443±0.171
	business unit	0.998±0.002	0.397±0.149
	search	0.994±0.011	0.439±0.140
	mobile	0.991±0.007	0.572±0.079
	media	0.997±0.002	0.500±0.051
	software	0.580±0.144	0.572±0.154
	mobile phone	0.998±0.003	0.617±0.093
	investment/inv.	0.683±0.268	0.553±0.256
	network	0.991±0.011	0.619±0.048
PR	internet	0.777±0.110	0.735±0.145
	e_mail	0.802±0.160	0.761±0.154
	business unit	0.702±0.104	0.857±0.022
	search	0.822±0.135	0.736±0.120
	mobile	0.946±0.067	0.633±0.054
	media	0.894±0.120	0.633±0.123
	software	0.493±0.195	0.728±0.178
	mobile phone	0.963±0.068	0.667±0.094
	investment/inv.	0.485±0.191	0.763±0.138
	network	0.993±0.010	0.619±0.048
recall	internet	0.745±0.120	0.745±0.133
	e_mail	0.722±0.165	0.789±0.116
	business unit	0.635±0.053	0.888±0.051
	search	0.709±0.177	0.807±0.113
	mobile	0.440±0.086	0.824±0.088
	media	0.536±0.154	0.825±0.095
	software	0.408±0.205	0.793±0.128
	mobile phone	0.392±0.039	0.919±0.035
	investment/inv.	0.405±0.077	0.797±0.110
	network	0.420±0.092	0.849±0.079
sR3	internet	1.000±0.000	0.075±0.033
	e_mail	1.000±0.000	0.042±0.028
	business unit	1.000±0.000	0.344±0.015
	search	1.000±0.000	0.364±0.016
	mobile	0.999±0.001	0.148±0.218
	media	1.000±0.000	0.477±0.021
	software	1.000±0.000	0.020±0.000
	mobile phone	1.000±0.000	0.582±0.026
	investment/inv.	0.999±0.001	0.099±0.092
	network	0.999±0.002	0.472±0.216

Measure	*conceptId*	*avg prec.*	*avg recall*
χ^2	art.int.	0.759±0.228	0.179±0.130
	online c.	0.615±0.196	0.217±0.174
	manufacturer	0.625±0.367	0.186±0.188
	wlan	0.836±0.202	0.212±0.153
	market c.	0.611±0.256	0.186±0.236
	cryptography	0.956±0.125	0.117±0.102
	mobile p.	0.892±0.142	0.133±0.061
	online g.	0.800±0.097	0.305±0.026
	know. man.	0.805±0.284	0.132±0.208
	price earn.r.	0.690±0.237	0.291±0.214
IG	art.int.	0.577±0.327	0.210±0.120
	online c.	0.548±0.217	0.226±0.170
	manufacturer	0.170±0.295	0.488±0.196
	wlan	0.533±0.350	0.320±0.178
	market c.	0.137±0.303	0.704±0.051
	cryptography	0.441±0.416	0.260±0.132
	mobile p.	0.487±0.404	0.225±0.126
	online g.	0.723±0.221	0.347±0.104
	know. man.	0.338±0.419	0.273±0.200
	price earn.r.	0.401±0.364	0.579±0.267
oddsRatio	art.int.	0.783±0.071	0.073±0.121
	online c.	0.829±0.044	0.064±0.053
	manufacturer	0.861±0.029	0.024±0.007
	wlan	0.933±0.016	0.104±0.137
	market c.	0.869±0.031	0.031±0.039
	cryptography	0.958±0.020	0.043±0.016
	mobile p.	0.652±0.093	0.121±0.031
	online g.	0.943±0.028	0.136±0.095
	know. man.	0.787±0.034	0.024±0.014
	price earn.r.	0.867±0.032	0.068±0.094
odds2	art.int.	0.005±0.001	0.510±0.149
	online c.	0.008±0.002	0.709±0.145
	manufacturer	0.016±0.011	0.840±0.103
	wlan	0.024±0.011	0.715±0.118
	market c.	0.026±0.004	0.777±0.024
	cryptography	0.028±0.005	0.830±0.117
	mobile p.	0.001±0.000	0.717±0.131
	online g.	0.018±0.003	0.796±0.107
	know. man.	0.011±0.003	0.512±0.145
	price earn.r.	0.021±0.005	0.879±0.041
PR	art.int.	0.759±0.228	0.179±0.130
	online c.	0.615±0.196	0.217±0.174
	manufacturer	0.542±0.366	0.204±0.181
	wlan	0.808±0.196	0.216±0.150
	market c.	0.611±0.256	0.186±0.236
	cryptography	0.956±0.125	0.117±0.102
	mobile p.	0.842±0.182	0.142±0.060
	online g.	0.800±0.097	0.305±0.026
	know. man.	0.805±0.284	0.132±0.208
	price earn.r.	0.690±0.237	0.291±0.214
recall	art.int.	0.003±0.000	0.815±0.089
	online c.	0.005±0.001	0.854±0.105
	manufacturer	0.009±0.001	0.949±0.026
	wlan	0.016±0.001	0.893±0.048
	market c.	0.112±0.312	0.887±0.064
	cryptography	0.023±0.003	0.954±0.017
	mobile p.	0.001±0.000	0.883±0.081
	online g.	0.015±0.005	0.887±0.076
	know. man.	0.006±0.001	0.776±0.096
	price earn.r.	0.012±0.002	0.972±0.016
sR3	art.int.	0.998±0.007	0.086±0.126
	online c.	0.943±0.181	0.019±0.010
	manufacturer	1.000±0.000	0.075±0.114
	wlan	1.000±0.000	0.081±0.022
	market c.	1.000±0.000	0.092±0.264
	cryptography	1.000±0.000	0.107±0.107
	mobile p.	0.875±0.270	0.087±0.080
	online g.	0.995±0.011	0.117±0.035
	know. man.	0.892±0.249	0.093±0.218
	price earn.r.	0.908±0.262	0.099±0.250

Table 3: This shows the average precision and recall over the top ten terms for the ten largest concepts (left) and our arbitrarily chosen ten small concepts (right)

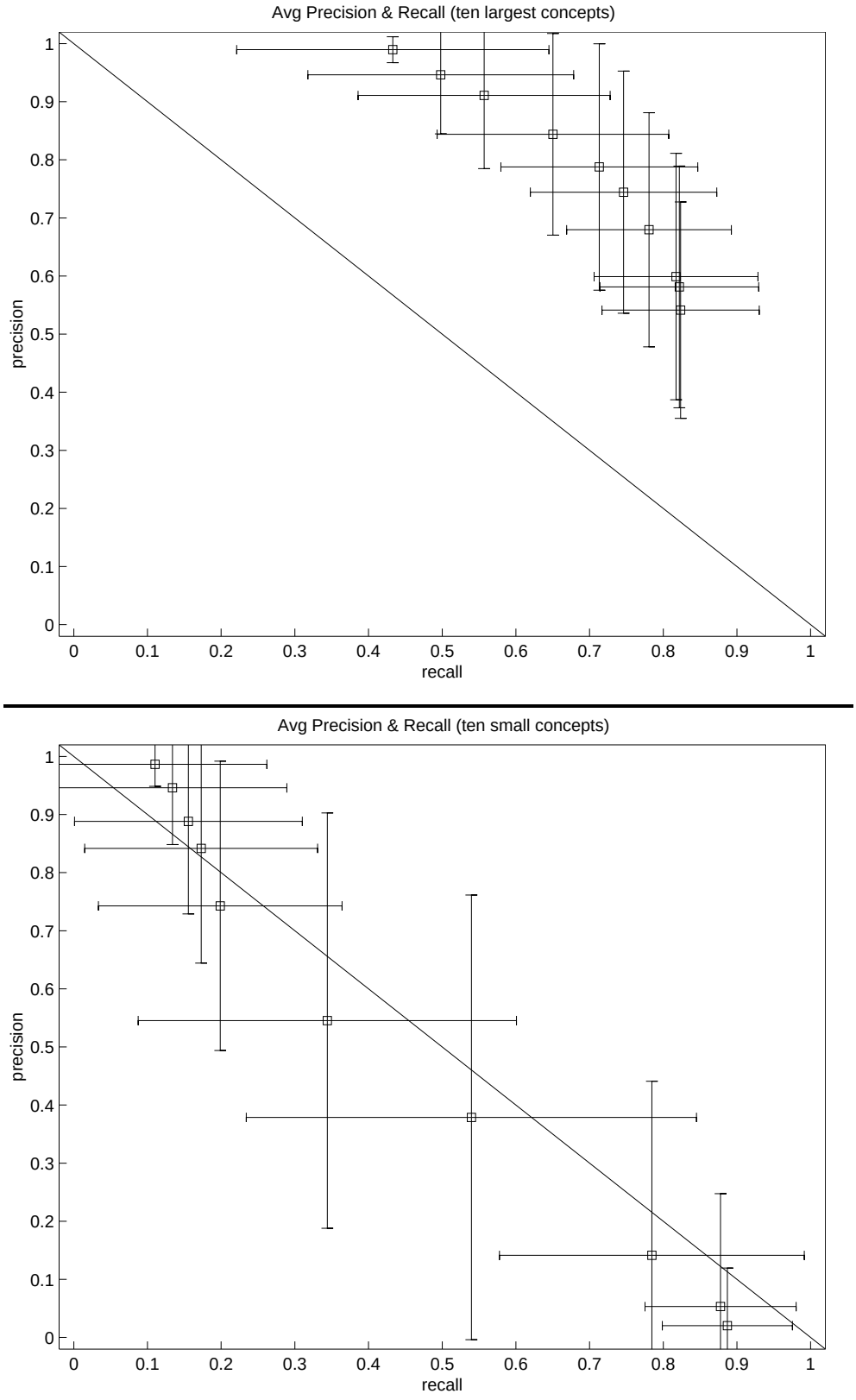


Figure 3: These plots show the average precision and recall $\pm$ standard deviation for the ten largest concepts (top) and our ten small concepts (bottom). The measures used here are $prec^x * recall^{2-x}$ where $0 \leq x < 2$ and steps of 0.2 were used, yielding ten points. $x = 0 \equiv$ *recall*, $x = 1 \equiv$ *PR*.

cept etc. can be instantly computed, thus allowing to give the user instant feedback during ontology editing.

Another idea would be to reduce the vocabulary to more manageable level – e.g. removing those terms which appear in less than ten documents would reduce the number of terms by an order of magnitude.

7 Conclusion

As we have seen, it is in principle possible to characterize ontology nodes by single words even in the presence of multi-word patterns. The results are somewhat promising, but in need of improvement. Our research would benefit from the following, roughly in order of importance.

- Reducing the terms by an order of magnitude. E.g. removing those terms which appear in less than 0.02% of all documents would be a feasible option.
- A Full-Text Index which scans for words and not for substrings would enable a more precise estimation of recall and precision.
- Precomputed contingency tables for each term and concept definition – or a way to compute these tables fast enough for realtime feedback.
- For people starting new ontologies with single word patterns, precomputed contingency tables for each combination of two terms would also be useful to approximate both multi-word patterns and multiple single word patterns. However, the memory requirements make this quite unfeasible – e.g. even for just 40,000 terms this would mean $6 * (40,000)^2 = 9.6 * 10^9$ values 40 Gigabytes memory consumption.

Acknowledgements

We want to thank Reinhard Schwab for optimizing our experimental java code and for giving valuable hints concerning runtime improvements. We also want to thank an anonymous colleague for being on vacation, so we could use his much faster computer for our evaluation of results.

References

Mladenic, D., (1998) Feature subset selection in text-learning, Proceedings of 10th European Conference on Machine Learning, 1998.

van Rijsbergen, C.J., Harper, D.J., Porter, M.F., The selection of good search terms, *Information Processing and Management*, 17, pp.77–91, 1981.

Yang, Y., Pedersen J.P. (1997). A Comparative Study on Feature Selection in Text Categorization Proceedings of the Fourteenth International Conference on Machine Learning (ICML'97), pp.412-420.

8 Appendix

conceptId	χ^2	*IG*	*oddsRatio*	*odds2*	*PR*	*recall*	*sR3*
	internet	**internet**	lebensversicherers	**internet**	**internet**	in	dispatch
	com	com	mamax	**web**	in	2001	newshilight
	web	**web**	mysimon	mysimon	2001	an	catchup
	it	it	techrepublic	techrepublic	com	**internet**	newslinksmall
internet	contact	contact	gamespot	gamespot	an	1	boldlink
	from	from	shopper	corrections	1	com	storyurl
	copyright	copyright	corrections	comparisons	online	at	openmpwin
	privacy	privacy	aktivierungsschlüssel	contact	jobs	jobs	bigopen
	about	for	mediadaten	sites	at	online	scrollable
	sites	about	bandwidth	reviews	it	die	gatrixx
	mail	**mail**	9pt	**mail**	**mail**	in	boldlink
	version	version	fußballverein	kontakt	version	2001	companyname
	e	td	lebensversicherers	version	jobs	an	smalltext
	jobs	**e**	mamax	hoppenstedt	**e**	jobs	mediumbold
e_mail	td	jobs	hoppenstedt	**email**	in	**e**	7pt
	kontakt	kontakt	marketperformer	leserbrief	2001	version	smalltext2
	archiv	archiv	javaenabled	drucken	news	news	newmediasales
	latest	color	javaon	verschicken	kontakt	**mail**	stammzellen
	tools	latest	gfhshplugintargetversion	archiv	an	1	ft26xx3044x11
	color	tools	gfhshpluginname	kurssuche	archiv	die	ft29xx3144xx1
	unternehmen	**unternehmen**	leserbriefe	**unternehmen**	**unternehmen**	2001	börsenticker
	weitere	weitere	**unternehmen**	leserbriefe	der	in	hoppenstedt
	kapital	contact	creditreform	creditreform	die	die	analystenstimmen
	ftd	das	stichwortsuche	stichwortsuche	und	der	munzinger
business unit	partnersites	für	icra	graumarkt	von	und	partnersites
	writelayersn	von	graumarkt	ftd	das	von	leserbriefe
	writerecherchen	mit	marketperformer	guided	für	mit	sitemap
	writezeitungn	kontakt	mdax	persönliches	mit	für	logout
	writemultichanneln	der	lbbw	zeitungs	den	den	menusactive
	writeregistrierenn	und	**unternehmens**	neuemissionen	im	an	writelayersn
	tools	news	leserbriefe	leserbriefe	news	in	börsenticker
	latest	tools	partnersites	partnersites	tools	2001	hoppenstedt
	news	latest	research	latest	latest	an	munzinger
	pda	pda	dispatch	analystenstimmen	jobs	news	stichwortsuche
search	times	mail	analystenstimmen	guided	e	jobs	sitemap
	mail	jobs	icra	graumarkt	mail	e	logout
	home	home	guided	recherche	pda	us	menusactive
	edition	times	graumarkt	tools	home	tools	writelayersn
	jobs	e	marketperformer	creditreform	in	latest	writerecherchen
	fi nancial	td	creditreform	topics	service	1	writezeitungn
	mobile	**mobile**	topics	**mobile**	**mobile**	in	dispatch
	policy	policy	**wireless**	topics	policy	2001	emazing
	headlines	headlines	5s	**wireless**	headlines	a	writeexchangelink
	pt	pt	**mobile**	headlines	pt	us	metricom
mobile	featured	search	catchup	pt	featured	com	convertedsymbol
	1995	featured	openwindow	featured	search	top	nextcard
	topics	services	unescape	policy	1995	at	**wireless**
	search	1995	getsurveyfi le	corrections	services	an	xis
	wireless	privacy	surveyfi le	mysimon	topics	s	topics
	services	about	pdt	gamespot	privacy	new	aprs
	media	**media**	leserbriefe	leserbriefe	**media**	in	börsenticker
	medien	**medien**	analystenstimmen	**medien**	**medien**	2001	hoppenstedt
	edition	edition	stichwortsuche	analystenstimmen	edition	an	munzinger
	times	times	graumarkt	stichwortsuche	times	**media**	partnersites
media	true	true	marketperformer	graumarkt	true	news	sitemap
	partnersites	überblick	creditreform	creditreform	version	version	logout
	writelayersn	partnersites	realmedia	guided	home	jobs	menusactive
	writerecherchen	writelayersn	redir	zeitungs	überblick	document	writelayersn
	writezeitungn	writerecherchen	guided	ftd	partnersites	write	writerecherchen
	writemultichanneln	writezeitungn	useragent	persönliches	writelayersn	die	writezeitungn
	software	**software**	**software**	**software**	**software**	in	geldverwalter
	microsoft	microsoft	untermieter	microsoft	microsoft	2001	witwenrente
	pc	pc	klingeling	web	web	**software**	sponsort
	web	web	fi bel	pc	pc	an	hilf
software	desktop	desktop	bauzulieferer	it	it	at	freudentänze
	product	product	malade	services	in	new	papiertiger
	this	it	papierchen	this	2001	us	fondspolice
	sites	this	glitzernde	to	com	com	goldies
	computer	computer	lpez	tech	a	1	erdrückende
	reserved	services	schlafmützen	computer	for	a	konjunkturgewitter
	handy	**handy**	kurssuche	kurssuche	**handy**	in	hoppenstedt
	stellenmarkt	stellenmarkt	börsenticker	börsenticker	stellenmarkt	2001	munzinger
	branchen	branchen	leserbriefe	leserbriefe	branchen	die	sitemap
	sms	sms	partnersites	partnersites	sms	der	logout
mobile phone	wap	abonnieren	analystenstimmen	analystenstimmen	wap	und	menusactive
	writelayersn	wap	stichwortsuche	**handy**	abonnieren	das	writelayersn
	writerecherchen	persönliches	graumarkt	stichwortsuche	writelayersn	an	writerecherchen
	writezeitungn	writelayersn	creditreform	graumarkt	writerecherchen	für	writezeitungn
	writemultichanneln	writerecherchen	marketperformer	creditreform	writezeitungn	mit	writemultichanneln
	writeregistrierenn	writezeitungn	**handy**	guided	writemultichanneln	den	writeregistrierenn
	investor	**investor**	dispatch	**investor**	**investor**	in	9ca380
	cfd6b2	a	**investment**	dispatch	a	2001	cfd6b2
	9ca380	line	todays	a	top	an	wirtschaftit
	rundschau	top	personalia	rundschau	line	a	dispatch
investment/investor	dispatch	new	**investor**	line	new	new	tandard
	a	s	lebhafte	top	s	top	**investment**
	line	context	drahtseilakt	**investment**	in	at	maxi
	context	cfd6b2	reichmacher	s	2001	us	geldverwalter
	weekly	9ca380	sponsort	new	us	s	schwalben
	starting	http	untermieter	time	time	line	witwenrente
	networks	**networks**	corrections	corrections	**networks**	2001	mysimon
	hot	hot	techrepublic	techrepublic	hot	in	dispatch
	mysimon	featured	gamespot	gamespot	reserved	us	catchup
	corrections	reserved	comparisons	**networks**	mysimon	news	techrepublic
network	comparisons	topics	**networks**	comparisons	comparisons	com	gamespot
	techrepublic	send	reviews	reviews	corrections	a	corrections
	gamespot	comparisons	**network**	reserved	techrepublic	s	**networks**
	reserved	mysimon	kary	send	gamespot	all	comparisons
	reviews	reviews	heavenly	zdnet	reviews	for	5s
	send	corrections	**netzwerk**	hot	send	new	**network**

Table 4: This table shows the top ten terms according to each measure for the largest ten concepts. The first (uppermost) term has rank one, the next lower one has rank two and so on.

conceptId	χ^2	*IG*	*oddsRatio*	*odds2*	*PR*	*recall*	*sR3*
artifi cial intelligence	**ai**	**ai**	**ai**	man	**ai**	2001	**ki**
	intelligenz	**intelligenz**	ziegelindustrie	height	**intelligenz**	in	**ai**
	ki	**artificial**	wunderwaffe	ist	**ki**	der	trappl
	artificial	**künstliche**	arbeitssklaven	eine	**artificial**	die	dunietz
	künstliche	**ki**	crossbar	kann	**künstliche**	und	hutchens
	seminarvorträge	**intelligence**	turing	können	seminarvorträge	von	verbmobil
	kampfroboter	seminarvorträge	milky	bin	kampfroboter	den	seminarvorträge
	wunderwaffe	privatstiftungen	descendant	online	wunderwaffe	das	goren
	arbeitssklaven	**künstlicher**	fbunts	welt	arbeitssklaven	für	treister
	privatstiftungen	wölfl	blinding	werden	privatstiftungen	mit	dfki
online community	blackbox	**community**	oberschalen	**internet**	blackbox	2001	musicity
	fuck	fuck	kemco	**online**	fuck	in	konop
	zwischenablagen	blackbox	pagern	web	zwischenablagen	**internet**	pocketmail
	playstations	playstations	dreiste	software	playstations	**online**	vitos
	handtasche	zumindestens	tabellenkalkulationen	microsoft	handtasche	an	mcnet
	elektron	mond	blackbox	pc	elektron	at	phonoverband
	zumindestens	schett	heimkehrende	net	**community**	1	markle
	community	zwischenablagen	sputnik7	at	zumindestens	us	wissenszentren
	mond	schrott	zwischenablagen	on	mond	com	thielen
	minigehäuse	handtasche	faxnachrichten	5	minigehäuse	e	mcowen
manufacturer	**handyhersteller**	**handyhersteller**	z100	**handy**	**handyhersteller**	2001	**handyhersteller**
	mobiltelefonhersteller	**handys**	bistum	als	**mobiltelefonhersteller**	in	**mobiltelefonhersteller**
	sendo	nokia	netzbau	für	sendo	der	**handyherstellern**
	handyherstellern	**hersteller**	farbiges	markt	**handyherstellern**	die	deighton
	mangment	ericsson	fi m	document	**mobiltelefone**	für	pretec
	mobiltelefone	**mobiltelefone**	jussi	das	**handys**	und	duftchip
	zusammenschaltung	**handy**	mobiltelefongeschäft	nokia	mangment	das	adventskalender
	aktienverkauf	mobilfunk	telefonino	bei	zusammenschaltung	mit	würzbug
	duftchip	gprs	abbauten	mit	aktienverkauf	von	besinnlichen
	handys	fi nnische	typografi e	den	fi nnische	den	wohlgeruch
wlan	**wlan**	**wlan**	**wlan**	**wireless**	**wlan**	2001	bankcomputer
	lan	**lan**	computersäulen	parent	**lan**	in	kummernummer
	802	802	döw	neue	802	die	kummermail
	11a	11a	überwachungsservice	microsoft	11a	der	netstumbler
	bankcomputer	**wireless**	methangas	wird	bankcomputer	und	radau
	grafi kchips	grafi kchips	funklösung	für	grafi kchips	den	**wlans**
	subnotebook	**lans**	internetexplorer	den	subnotebook	mit	prozessorentechnologie
	kummernummer	bluetooth	computerschrott	var	kummernummer	für	informatikbegriffe
	kummermail	notebooks	mönche	mit	kummermail	von	betastadium
	radau	bankcomputer	mov	ist	komma	das	kutten
market capitalisation	**marktkapitalisierung**	**marktkapitalisierung**	**capitalization**	fi nancial	**marktkapitalisierung**	2001	**marktkapitalisierung**
	capitalization	börsen	angeschwollen	latest	**capitalization**	in	kursänd
	rangliste	neuemissionen	umschreibung	tools	rangliste	an	behrent
	zwischenbericht	persönliches	oica	aktien	zwischenbericht	die	punktabzug
	autokonzerne	sitemap	wertpapiermärkten	times	autokonzerne	der	furse
	oica	logout	schwieg	euro	oica	und	kleiman
	vda	munzinger	zeichnungsgewinn	7	vda	mit	goy
	monatszahlen	writelayersn	auswahlkriterium	unternehmen	reinking	den	skyy
	reinking	writerecherchen	ibiza	pda	monatszahlen	für	thelemann
	börsenumsatz	writezeitungn	bewertungs	home	börsenumsatz	**marktkapitalisierung**	indexrevision
cryptography	**verschlüsselung**	**verschlüsselung**	animator	internet	**verschlüsselung**	die	**verschlüsselung**
	verschlüsseln	**verschlüsseln**	freundet	ist	**verschlüsseln**	in	**verschlüsseln**
	verschlüsselte	netzpolitik	maßvoll	werden	**verschlüsselte**	und	**verschlüsselt**
	verschlüsselt	webstandard	virenliste	auf	**verschlüsselt**	der	**verschlüsselte**
	verschlüsselungs	hacker	winlinux	date	**verschlüsselungs**	2001	**verschlüsselungs**
	verschlüsseltem	mails	internetüberwachung	für	nsa	für	**verschlüsseltem**
	nsa	fi rewall	maschen	oder	**verschlüsseltem**	von	**kryptographie**
	animator	**verschlüsselte**	inkompatiblen	das	animator	den	kompostierbaren
	freundet	nsa	musikfan	den	freundet	das	**verschlüsselungssoftware**
	kompostierbaren	umfrage	mobilfunkvertrag	vor	kompostierbaren	auf	bildschirmschoners
mobile portal	**mobilfunkportal**	**mobilfunkportal**	unterhaltsames	**handy**	**mobilfunkportal**	in	**mobilfunkportal**
	contentbereich	jamba	gevay	com	contentbereich	an	microphones
	betriebssystemunterstützung	travelchannel	internetfernsehen	de	betriebssystemunterstützung	2001	contentbereich
	travelchannel	kundrun	travelchannel	**mobile**	travelchannel	com	netsize
	unterhaltsames	gerätes	skywire	soll	unterhaltsames	der	betriebssystemunterstützung
	gevay	wirtschaftspresse	handyabsatzes	einer	gevay	den	targus
	internetfernsehen	stadtplan	playerlösung	internet	internetfernsehen	und	datenfähigkeit
	datenfähigkeit	contentbereich	sugarman	i	datenfähigkeit	das	uninstalled
	communitysektor	betriebssystemunterstützung	kundrun	an	kundrun	die	chesnais
	umsatzsteigung	**portal**	faircar	2002	communitysektor	von	wichmann
online gaming	startopia	**games**	spielehits	**online**	unreal	in	**onlinespiele**
	unreal	unreal	undying	the	startopia	**online**	tekken
	xtreme	startopia	barkers	of	xtreme	2001	elektroschocks
	racer	xtreme	rennbahn	top	racer	the	flrtspielchen
	dungeon	dungeon	einköpfi ge	line	dungeon	top	biershooter
	eidos	racer	cybersex	a	eidos	a	nashorn
	siege	siege	interplay	microsoft	siege	at	fallschirmspringen
	quake	eidos	gehirne	4	quake	of	spielehits
	commandos	desperados	startopia	to	commandos	an	multimediabrille
	desperados	kabuto	xtreme	was	desperados	new	barkers
knowledge management	**km**	**km**	sensational	email	**km**	in	**km**
	wissensmanagement	**wissensmanagement**	cmb	online	**wissensmanagement**	2001	**wissensmanagement**
	karraker	kmart	karraker	com	karraker	1	artefact
	kmart	bluelight	ambulanten	dieser	kmart	at	x83
	richtfunkantenne	mart	jcp	version	richtfunkantenne	online	langlauf
	dissuaded	email	troger	meldungen	dissuaded	com	nbsp
	wmt	karraker	sprachsoftware	seite	wmt	an	artibrain
	sensational	**knowledge**	kommunikationsebene	wien	sensational	die	airlancer
	cmb	ansprechpartner	sprachlösungen	pte	cmb	und	audigy
	begriffswelt	druckerfreundliche	glasfaserleitung	produkte	begriffswelt	von	siriri
price earning ratio	**kgv**	**kgv**	gewinnwarnungsreigen	wir	**kgv**	2001	**kgv**
	aktienrückkäufe	rechtfertigen	emagine	aktien	aktienrückkäufe	in	frauenförderung
	rechtfertigen	aktienrückkäufe	aktienrückkäufe	unternehmen	rechtfertigen	die	kursänd
	sommerrally	lignum	irrglaube	uns	sommerrally	der	punktabzug
	lignum	sommerrally	stammzelltherapien	mehr	lignum	und	skyy
	brainlab	brainlab	profi tablere	zum	brainlab	für	steinemann
	winkt	fonds	gewinnhalbierung	aus	winkt	von	prozeß
	frauenförderung	**kurse**	bewertungs	über	frauenförderung	auf	gejamm
	gewinnhalbierung	aktien	hochkapitalisierten	heute	gewinnhalbierung	bei	emagine
	deepwater	börsen	rückschlagpotenzial	bei	deepwater	im	jurka

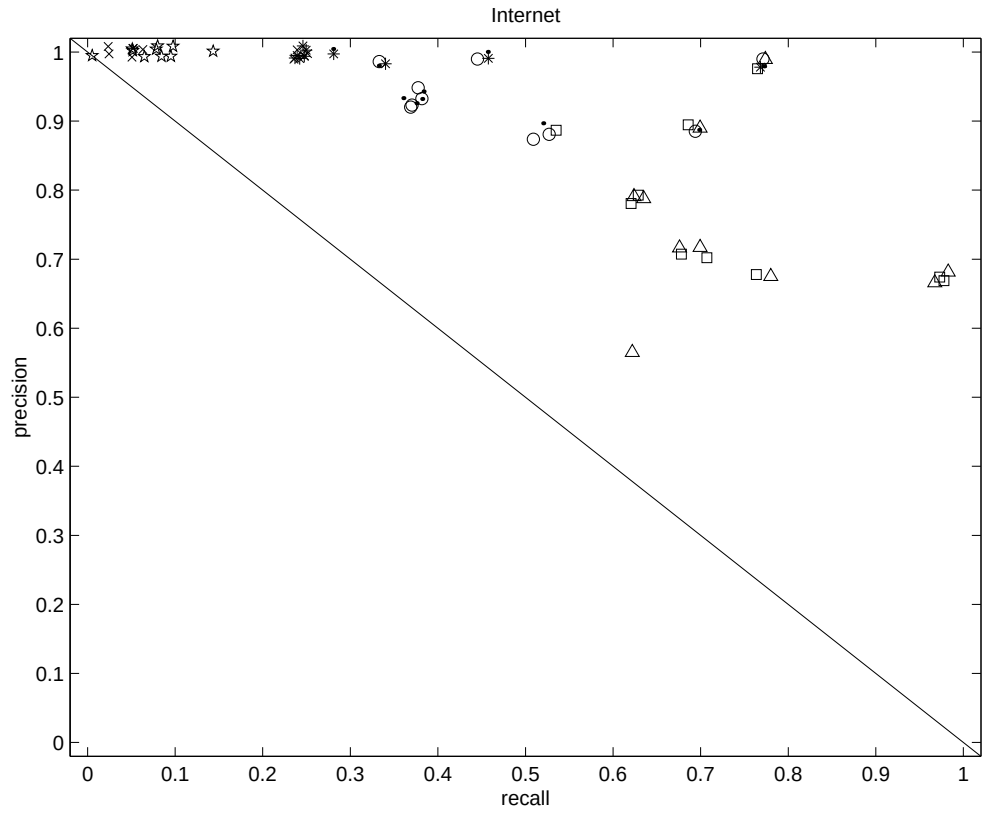
Table 5: This table shows the top ten terms according to each measure for our ten small concepts. The first (uppermost) term has rank one, the next lower one has rank two and so on.

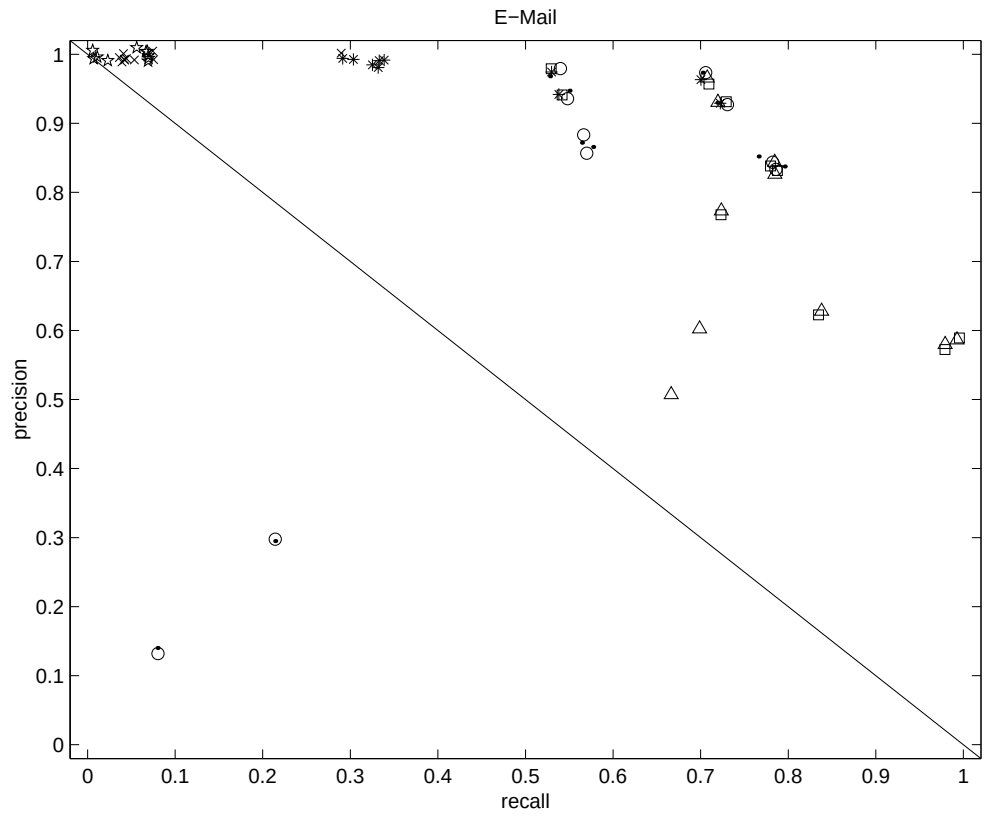
conceptId	χ^2	*IG*	*oddsRatio*	*odds2*	*PR*	*recall*	*sR3*
	internet	**internet**	mediadaten	**internet**	**internet**	2001	newshilight
	com	com	menubar	**web**	in	in	newslinksmall
	web	wirtschaftblatt	bookshop	contact	2001	**internet**	boldlink
	wirtschaftblatt	atx	lebensversicherers	com	com	1	gatrixx
internet	atx	ifg	mamax	menubar	1	an	storyurl
	ifg	easi	ck	mysimon	an	com	argnr
	easi	vidx	mysimon	techrepublic	for	for	getday
	vidx	währungen	techrepublic	gamespot	news	news	getmonth
	währungen	smi	susa	sites	new	document	getdate
	kursliste	kursliste	gamespot	nachrichtendienst	us	us	isnmonths
	version	version	fi ndobj	**email**	version	in	boldlink
	mail	**mail**	9pt	**mail**	**mail**	2001	companyname
	e	**e**	fußballverein	version	**e**	an	smalltext
	email	**email**	lebensversicherers	kontakt	in	1	mediumbold
e_mail	kontakt	td	mamax	**e**	2001	version	7pt
	td	wirtschaftblatt	hoppenstedt	printer	an	**e**	smalltext2
	jobs	ifg	javaenabled	hoppenstedt	jobs	com	thiskind
	media	easi	javaon	druckerfreundliche	news	us	uniscape
	home	vidx	gfhshplugintargetversion	aussender	us	news	newmediasales
	archiv	atx	gfhshpluginname	friendly	**email**	document	stammzellen
	unternehmen	**unternehmen**	leserbriefe	**unternehmen**	**unternehmen**	2001	börsenticker
	weitere	contact	creditreform	leserbriefe	der	in	hoppenstedt
	kontakt	weitere	stichwortsuche	creditreform	von	die	analystenstimmen
	kapital	das	icra	stichwortsuche	2001	der	munzinger
business unit	ftd	für	**unternehmen**	graumarkt	und	an	partnersites
	partnersites	von	börsengänge	ftd	die	und	sitemap
	creditreform	im	init	guided	in	von	logout
	writelayersn	den	68k	zeitungs	das	mit	menusactive
	writerecherchen	mit	**enterprise**	abonnenten	für	das	writelayersn
	writezeitungn	quotes	graumarkt	einrichten	mit	für	writerecherchen
	news	news	leserbriefe	topics	news	in	börsenticker
	tools	tools	partnersites	featured	tools	2001	hoppenstedt
	latest	latest	besucher	headlines	in	news	munzinger
	times	times	research	recherche	home	an	stichwortsuche
search	policy	policy	realmedia	leserbriefe	us	1	sitemap
	pt	pt	icra	partnersites	latest	us	logout
	headlines	edition	redir	pt	times	document	menusactive
	1995	1995	analystenstimmen	1995	service	for	writelayersn
	edition	headlines	dispatch	policy	2001	com	writerecherchen
	featured	**search**	erotikangebote	latest	7	write	writezeitungn
	mobile	**mobile**	topics	**mobile**	**mobile**	in	dispatch
	wireless	**wireless**	5s	**wireless**	**wireless**	2001	emazing
	policy	policy	**wireless**	topics	policy	for	writeexchangelink
	pt	pt	**mobile**	pt	pt	1	convertedsymbol
mobile	headlines	services	subscribe	headlines	services	us	xis
	1995	headlines	advertisement	featured	headlines	an	nextcard
	featured	1995	maximize	policy	tech	news	aprs
	services	tech	openwindow	1995	1995	com	compq
	search	search	unescape	magazine	search	a	symbol1
	magazine	featured	getsurveyfi le	related	its	s	symbol2
	media	**media**	leserbriefe	**medien**	**media**	in	börsenticker
	edition	edition	stichwortsuche	leserbriefe	edition	2001	hoppenstedt
	times	times	realmedia	**media**	times	**media**	munzinger
	medien	true	redir	edition	true	an	partnersites
media	true	**medien**	analystenstimmen	stichwortsuche	home	1	sitemap
	navigator	home	relevancy	msie	version	document	logout
	style	style	creditreform	analystenstimmen	7	news	menusactive
	home	navigator	useragent	creditreform	service	write	writelayersn
	msie	version	illustration	useragent	**medien**	us	writerecherchen
	überblick	7	rm	appversion	news	version	writezeitungn
	software	**software**	**software**	**software**	**software**	2001	tagesüberblick
	microsoft	microsoft	papierchen	microsoft	microsoft	in	webpromotion
	computer	web	klingeling	it	in	**software**	newscasts
	web	it	bauzulieferer	web	it	an	newslinx
software	pc	computer	malade	computer	2001	1	internetational
	product	pc	glitzernde	pc	for	com	geldverwalter
	it	product	lpez	for	internet	for	witwenrente
	desktop	for	schlafmützen	more	com	internet	sponsort
	this	this	wahlzeit	services	web	us	hilf
	mobile	mobile	wurlitzer	this	us	news	freudentänze
	handy	**handy**	börsenticker	**handy**	**handy**	in	hoppenstedt
	stellenmarkt	sms	kurssuche	börsenticker	sms	2001	munzinger
	sms	stellenmarkt	leserbriefe	kurssuche	stellenmarkt	die	sitemap
	branchen	branchen	partnersites	leserbriefe	branchen	der	logout
mobile phone	wap	wap	stichwortsuche	partnersites	wap	und	menusactive
	persönliches	technik	analystenstimmen	stichwortsuche	verschicken	an	writelayersn
	verschicken	verschicken	**handy**	analystenstimmen	persönliches	das	writerecherchen
	writelayersn	persönliches	creditreform	creditreform	technik	für	writezeitungn
	writerecherchen	writelayersn	icra	graumarkt	writelayersn	mit	writemultichanneln
	writezeitungn	writerecherchen	schlussbericht	stellenmarkt	writerecherchen	den	writeregistrierenn
	investor	**investor**	todays	**investor**	**investor**	in	9ca380
	investment	a	dispatch	**investment**	a	2001	cfd6b2
	a	line	**investment**	a	top	an	wirtschaftit
	dispatch	top	sponsort	top	line	a	tandard
investment/investor	context	**investment**	drahtseilakt	line	s	new	gilder
	line	s	klingeling	s	new	us	maxi
	weekly	time	bauzulieferer	new	time	top	geldverwalter
	investoren	tech	malade	time	in	1	schwalben
	starting	context	lpez	dispatch	us	s	witwenrente
	stay	new	moravec	tech	2001	at	glitzernde
	networks	**networks**	techrepublic	**networks**	**networks**	in	mysimon
	network	topics	corrections	techrepublic	**network**	2001	dispatch
	mysimon	inc	gamespot	corrections	reserved	us	catchup
	corrections	services	**networks**	gamespot	topics	for	5s
network	comparisons	this	**network**	**network**	mysimon	news	chasm
	techrepublic	today	comparisons	comparisons	comparisons	com	**netzwerken**
	gamespot	featured	todays	reviews	today	an	newsmakers
	reserved	policy	**netzwerk**	reserved	corrections	s	mariano
	reviews	**network**	eweek	zdnet	techrepublic	1	gwendolyn
	send	reserved	emachines	send	gamespot	a	bulletproofi ng

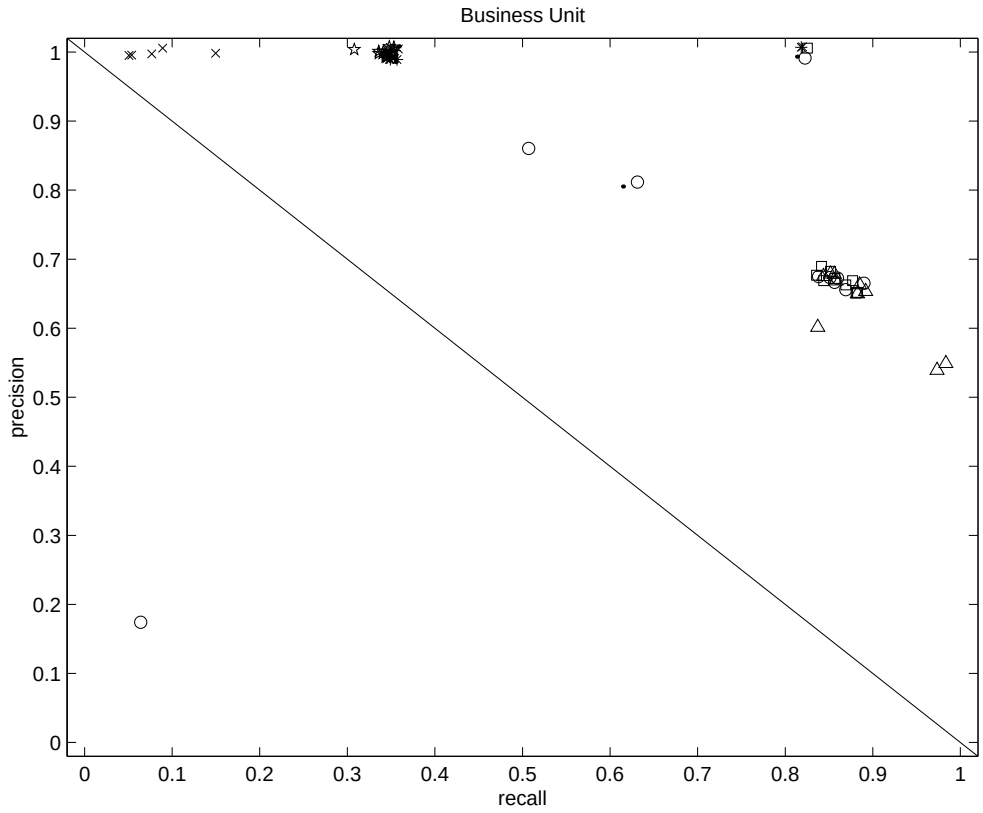
Table 6: This table shows the top ten terms according to each measure for the largest ten concepts when normalizing job size before combining the cotingency tables. The first (uppermost) term has rank one, the next lower one has rank two and so on.

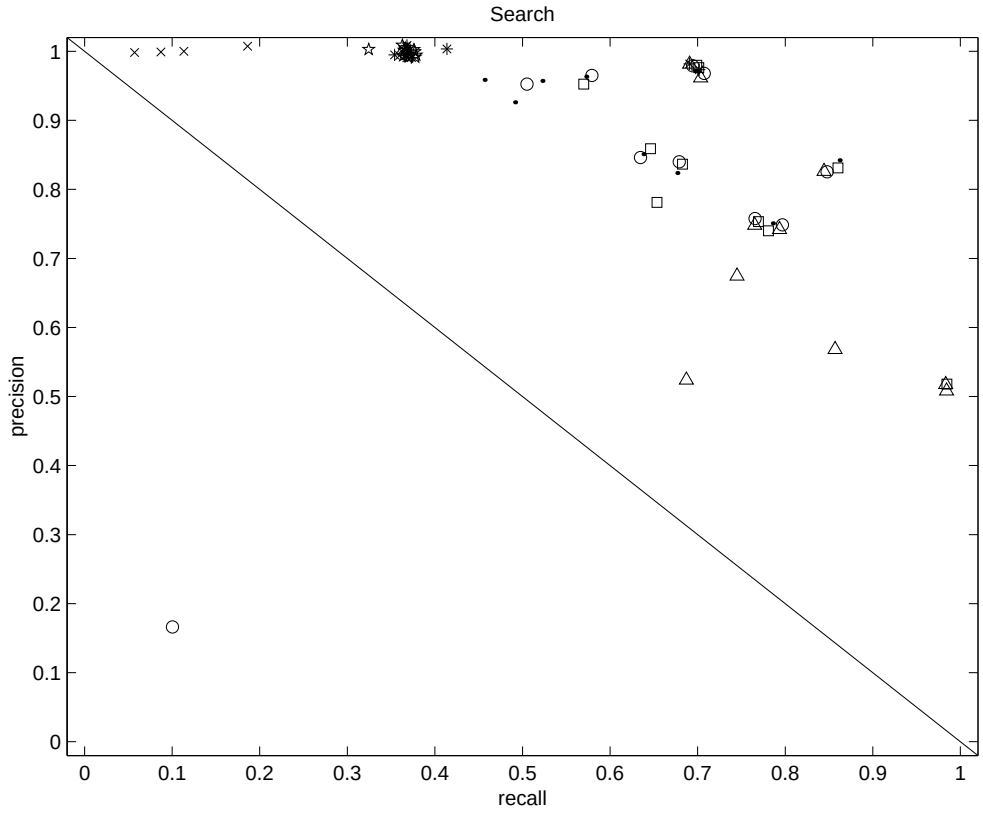
conceptId	χ^2	*IG*	*oddsRatio*	*odds2*	*PR*	*recall*	*sR3*
artificial intelligence	**ai**	**ai**	**ai**	eine	**ai**	2001	**ki**
	intelligenz	**intelligenz**	zeitaufwendig	ist	**intelligenz**	in	dunietz
	ki	**künstliche**	roboterentwicklung	von	**ki**	der	seminarvorträge
	seminarvorträge	**artificial**	monatsperformance	bin	seminarvorträge	1	hutchens
	künstliche	**intelligence**	ziegelindustrie	der	**künstliche**	die	goren
	artificial	**ki**	luxinvest	vor	**artificial**	und	treister
	künstlicher	seminarvorträge	anlageinstrumenten	welt	**künstlicher**	von	topicalnet
	privatstiftungen	privatstiftungen	nordinternet	nur	privatstiftungen	das	dfki
	suchtechnologie	**künstlicher**	turing	das	suchtechnologie	für	trappl
	dfki	zusammenfassung	inspizieren	kann	dfki	den	verbmobil
online community	pocketmail	**community**	kemco	**online**	pocketmail	2001	pocketmail
	kemco	pocketmail	faxnachrichten	**internet**	kemco	in	markle
	faxnachrichten	kemco	sputnik7	e	faxnachrichten	**online**	musicity
	oberschalen	faxnachrichten	rocksänger	software	oberschalen	**internet**	stabilisierungsanleihe
	pagern	oberschalen	oberschalen	web	pagern	an	mindestlohnes
	handlicher	pagern	pagern	pc	handlicher	1	konop
	nachstehen	handlicher	spieleproduzent	com	**community**	com	targus
	community	nachstehen	handlicher	microsoft	nachstehen	**community**	thumbpad
	auswechselbaren	auswechselbaren	dreiste	service	auswechselbaren	e	flrtmaster
	modisches	modisches	tabellenkalkulationen	new	modisches	at	vitos
manufacturer	**handyhersteller**	**handyhersteller**	bistum	**handy**	**handyhersteller**	2001	**handyhersteller**
	duftchip	**handys**	betriebsystems	für	duftchip	in	pretec
	pretec	nokia	textbasierten	das	pretec	die	duftchip
	adventskalender	**mobiltelefone**	farbiges	dem	adventskalender	der	adventskalender
	würzbug	**handy**	z100	ein	würzbug	und	würzbug
	besinnlichen	**mobiltelefon**	potentiell	markt	besinnlichen	für	besinnlichen
	wohlgeruch	**hersteller**	fi nanstidningen	hat	wohlgeruch	das	wohlgeruch
	bistum	klingeltöne	energiehunger	der	bistum	von	strahlungstests
	betriebsystems	gprs	simo	von	betriebsystems	mit	sommerprojekt
	sendo	mobilfunk	anzuhängen	die	sendo	an	nachwuchsforscher
wlan	**wlan**	**lan**	überwachungsservice	**wireless**	**wlan**	2001	auskunftssuchenden
	lan	**wlan**	computersäulen	neue	**lan**	in	patrone
	computersäulen	**wireless**	beobachtungs	parent	computersäulen	die	gaspatrone
	auskunftssuchenden	thiskind	wistron	eine	auskunftssuchenden	1	methangas
	patrone	fi ndobj	mov	internet	patrone	der	busbestellung
	gaspatrone	v4	stündiges	den	gaspatrone	und	protege
	überwachungsservice	mm	**wlan**	wird	überwachungsservice	mit	**wlans**
	beobachtungs	versenden	produktionsaktivitäten	für	beobachtungs	an	federgewicht
	methangas	multimedia	methangas	mit	methangas	den	moorhen
	gridpatrol	vereinigten	lebenslänglich	hat	gridpatrol	von	skycross
market capitalisation	**marktkapitalisierung**	**marktkapitalisierung**	angeschwollen	euro	**marktkapitalisierung**	2001	**marktkapitalisierung**
	capitalization	**capitalization**	oica	unternehmen	**capitalization**	in	splitt
	rangliste	börsen	**capitalization**	times	rangliste	an	kursänd
	zwischenbericht	börse	schwieg	7	zwischenbericht	die	behrent
	börsentagen	sitemap	jama	am	börsentagen	der	punktabzug
	autokonzerne	logout	retailgeschäfts	fi nancial	autokonzerne	und	furse
	oica	munzinger	sektorstrategie	tools	oica	**marktkapitalisierung**	kleiman
	vda	writelayersn	auswahlkriterium	vom	vda	von	goy
	börsenumsatz	writerecherchen	bewertungs	home	reinking	für	geldbrief
	reinking	writezeitungn	mehrheitsentscheidungen	aktien	börsenumsatz	mit	skyy
cryptography	**verschlüsselung**	**verschlüsselung**	bedienbaren	auf	**verschlüsselung**	die	**verschlüsselung**
	verschlüsselte	**verschlüsselte**	freihand	für	**verschlüsselte**	und	**verschlüsselte**
	verschlüsseln	technologie	mobilfunkvertrag	werden	**verschlüsseln**	der	**verschlüsselt**
	verschlüsselt	netz	inkompatiblen	ist	**verschlüsselt**	in	**verschlüsseln**
	bildmitteilung	stellt	netzstandards	den	inkompatiblen	für	**kryptographie**
	thumbboard	für	animator	das	bildmitteilung	den	thumbboard
	unterhaltungsmaschine	und	freundet	mit	thumbboard	mit	bildmitteilung
	inkompatiblen	alle	bereue	eine	unterhaltungsmaschine	auf	**verschlüsselungssoftware**
	bedienbaren	length	freihändige	und	bedienbaren	von	unterhaltungsmaschine
	freihand	auf	archos	von	freihand	das	**kryptografie**
mobile portal	**mobilfunkportal**	kabellos	unterhaltsames	**handy**	**mobilfunkportal**	in	**mobilfunkportal**
	contentbereich	pocketmail	gevay	**mobile**	contentbereich	an	targus
	betriebssystemunterstützung	kemco	internetfernsehen	i	betriebssystemunterstützung	2001	vpen
	zusatzservices	adapter	playerlösung	com	zusatzservices	com	thumbpad
	pocketmail	faxnachrichten	handyabsatzes	einer	pocketmail	1	microphones
	kemco	oberschalen	tant	de	kemco	der	netsize
	gerätepreise	pagern	zeichenerkennung	soll	gerätepreise	den	contentbereich
	unterhaltsames	handlicher	notbook	sich	unterhaltsames	und	betriebssystemunterstützung
	gevay	nachstehen	zusatzservices	an	gevay	das	uninstalled
	internetfernsehen	auswechselbaren	ideenaustausch	var	internetfernsehen	die	chesnais
online gaming	**gaming**	**gaming**	vanuatu	**online**	**gaming**	**online**	ussr
	unreal	**games**	spielehits	the	unreal	in	**onlinespiele**
	startopia	playstation	kommunikationsprogramms	microsoft	startopia	2001	tekken
	xtreme	**spiele**	undying	of	xtreme	the	elektroschocks
	racer	**game**	kampfpreis	top	racer	an	fanaticism
	dungeon	entertainment	virenprobleme	line	dungeon	if	gighahertz
	vanuatu	xbox	drittanbieter	4	vanuatu	top	twi
	aufrütteln	unreal	aufrütteln	a	aufrütteln	internet	flrtspielchen
	almanach	computerspiele	rennbahn	to	almanach	for	biershooter
	desperados	**online**	gehirne	entertainment	desperados	document	saboteure
knowledge management	**km**	**km**	karraker	email	**km**	in	**km**
	wissensmanagement	kmart	ambulanten	version	**wissensmanagement**	2001	**wissensmanagement**
	kmart	**wissensmanagement**	cmb	online	kmart	1	siriri
	karraker	bluelight	glasfaserleitung	11	karraker	com	langlauf
	ibi	mart	underestimating	com	ibi	online	sauter
	bluelight	retailing	lichtwellenleiter	13	bluelight	**km**	kph
	dissuaded	retailers	kernteam	at	dissuaded	an	hevs
	lichtwellenleiter	email	srdf	customer	lichtwellenleiter	at	b61
	underestimating	discount	asb	marketing	wmt	version	murzuk
	wmt	karraker	goggle	one	underestimating	11	lc4
price earning ratio	**kgv**	**kgv**	gewinnwarnungsreigen	bei	**kgv**	2001	**kgv**
	aktienrückkäufe	rechtfertigen	emagine	unternehmen	aktienrückkäufe	die	frauenförderung
	rechtfertigen	aktienrückkäufe	irrglaube	als	rechtfertigen	in	kursänd
	sommerrally	**kurse**	stammzelltherapien	sich	sommerrally	der	punktabzug
	lignum	lignum	profi tablere	aus	lignum	und	erstnotizausblick
	brainlab	fonds	schlussniveau	einem	brainlab	für	geldbrief
	frauenförderung	brainlab	aktienrückkäufe	zum	frauenförderung	von	primärmärkten
	gewinnhalbierung	graumarkt	bookrunner	über	winkt	mit	spwx
	winkt	sommerrally	gewinnhalbierung	ein	gewinnhalbierung	den	skyy
	trüber	aktien	bewertungs	für	trüber	auf	steinemann

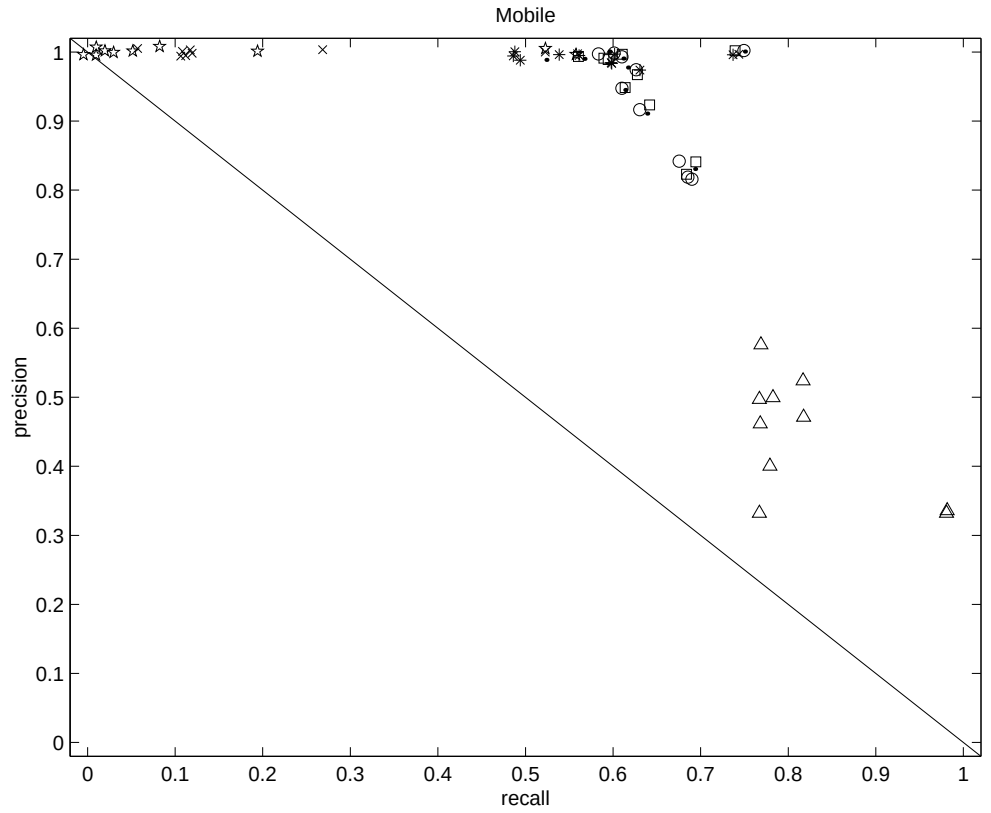
Table 7: This table shows our the top ten terms according to each measure for our ten small concepts when normalizing job size before combining the contingency tables. The first (uppermost) term has rank one, the next lower one has rank two and so on.

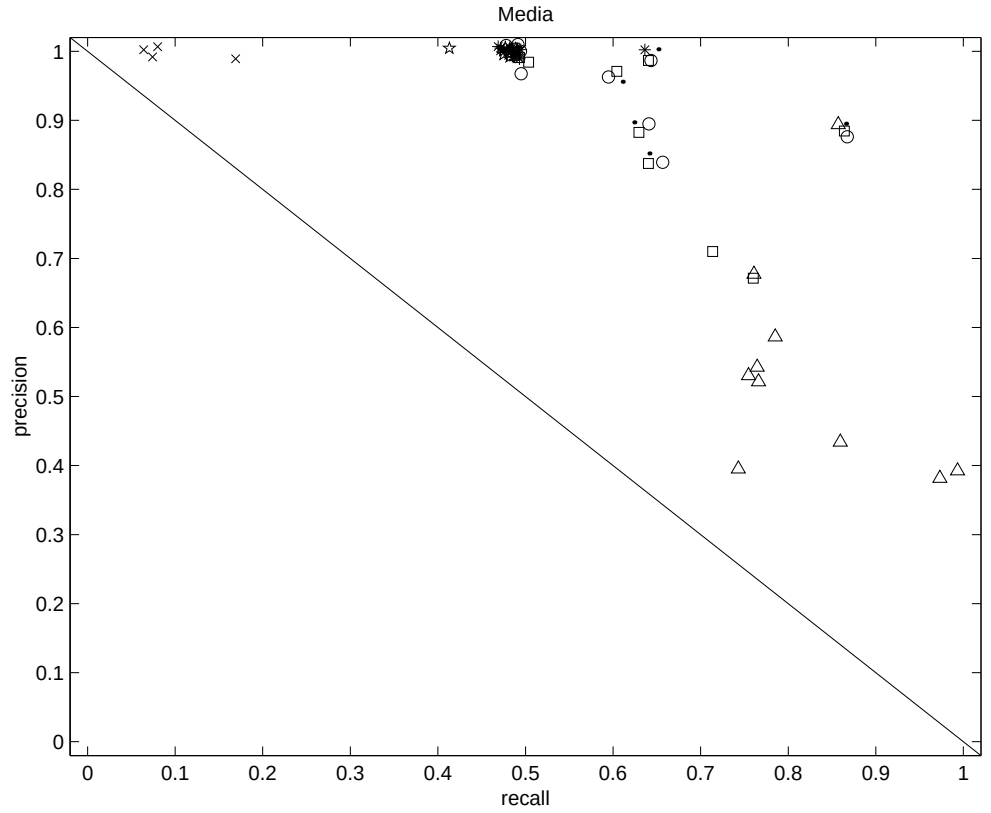

Internet
precision
recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

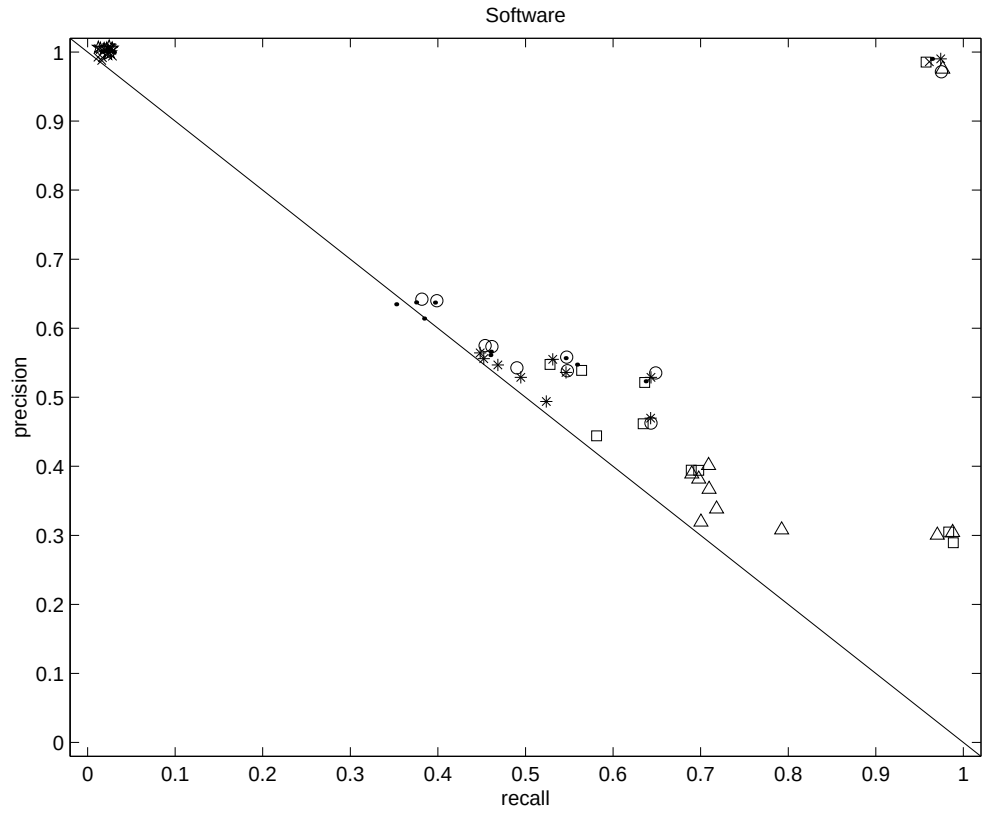

E−Mail
precision
recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

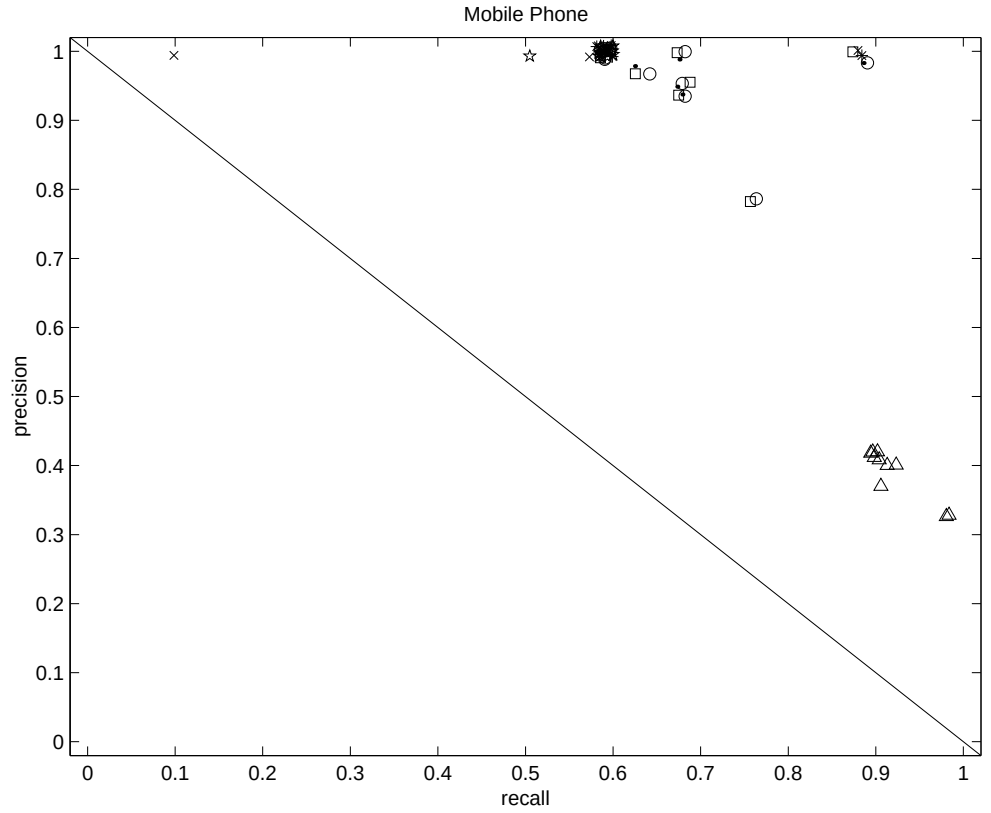

Business Unit
precision
recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

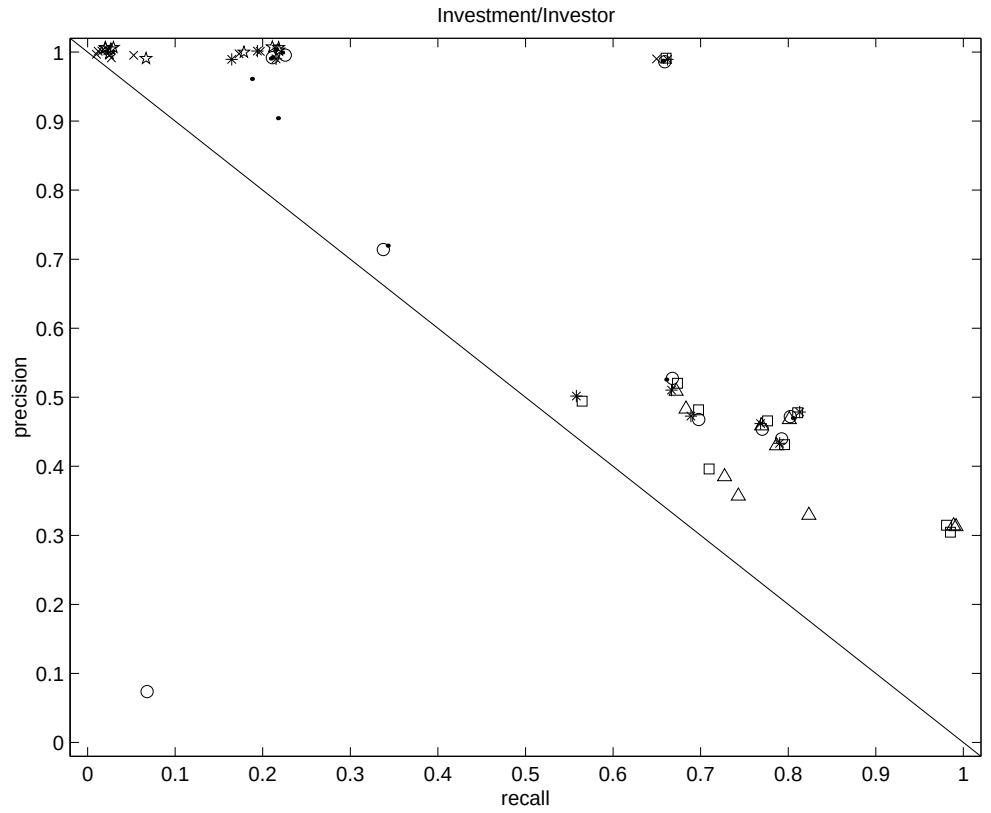

Search
precision
recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

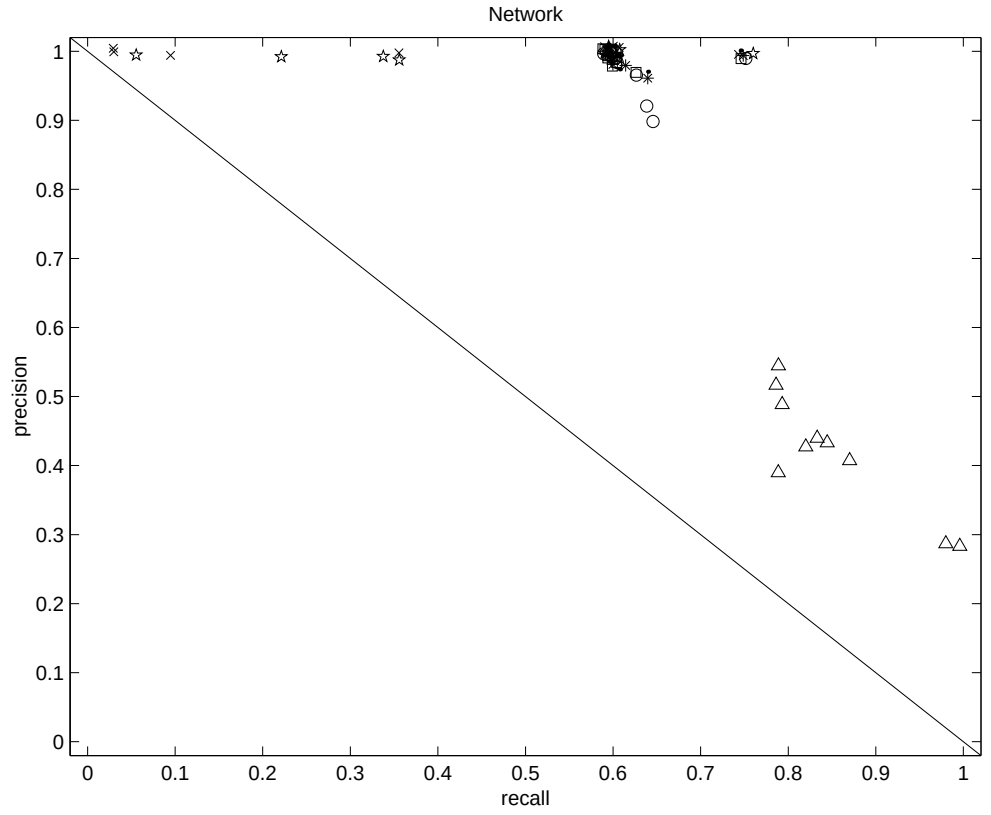

Mobile
precision
recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

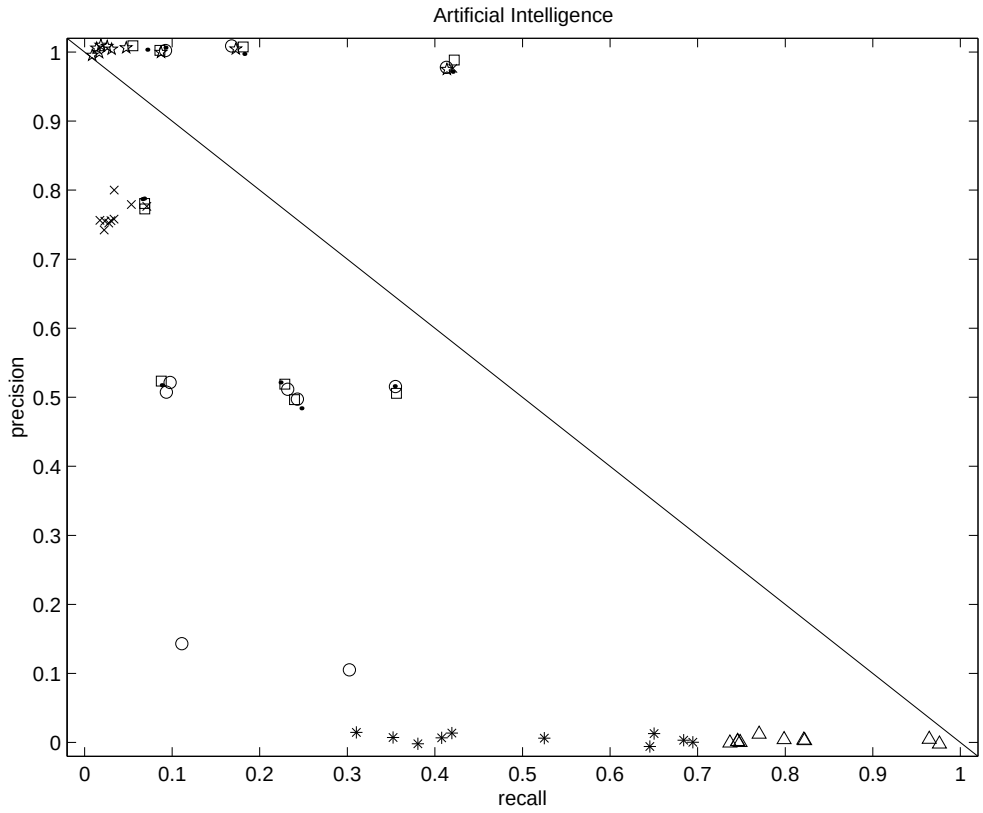

Media
precision
recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

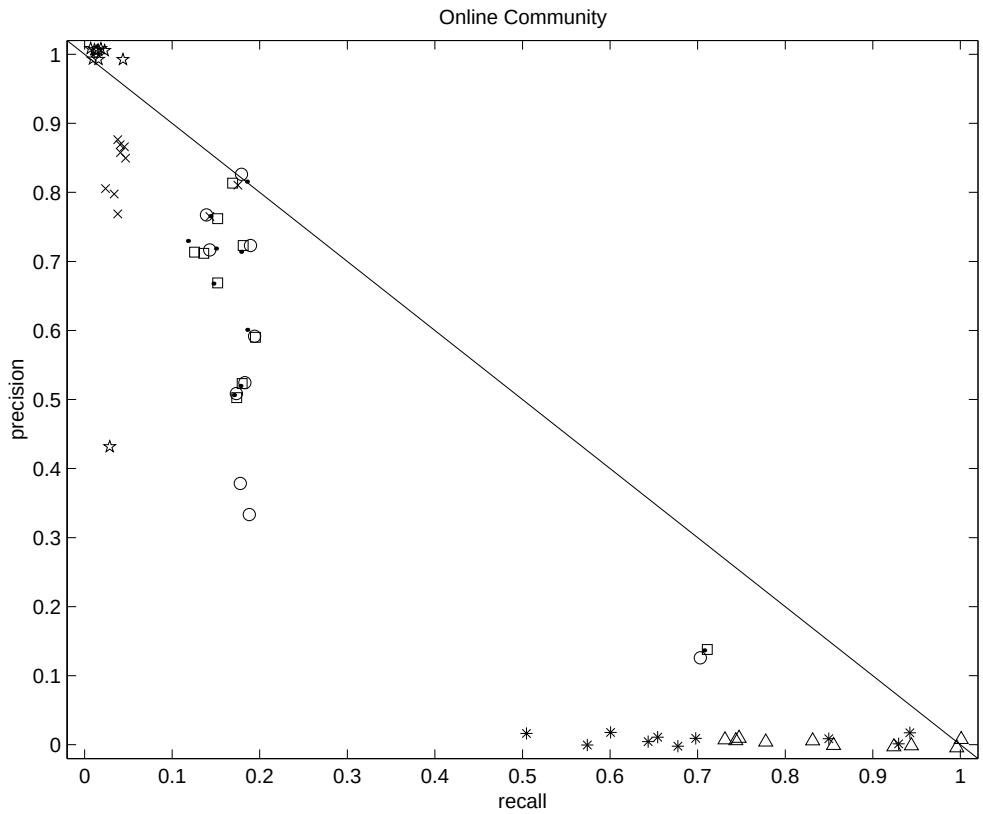

Software
precision
recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

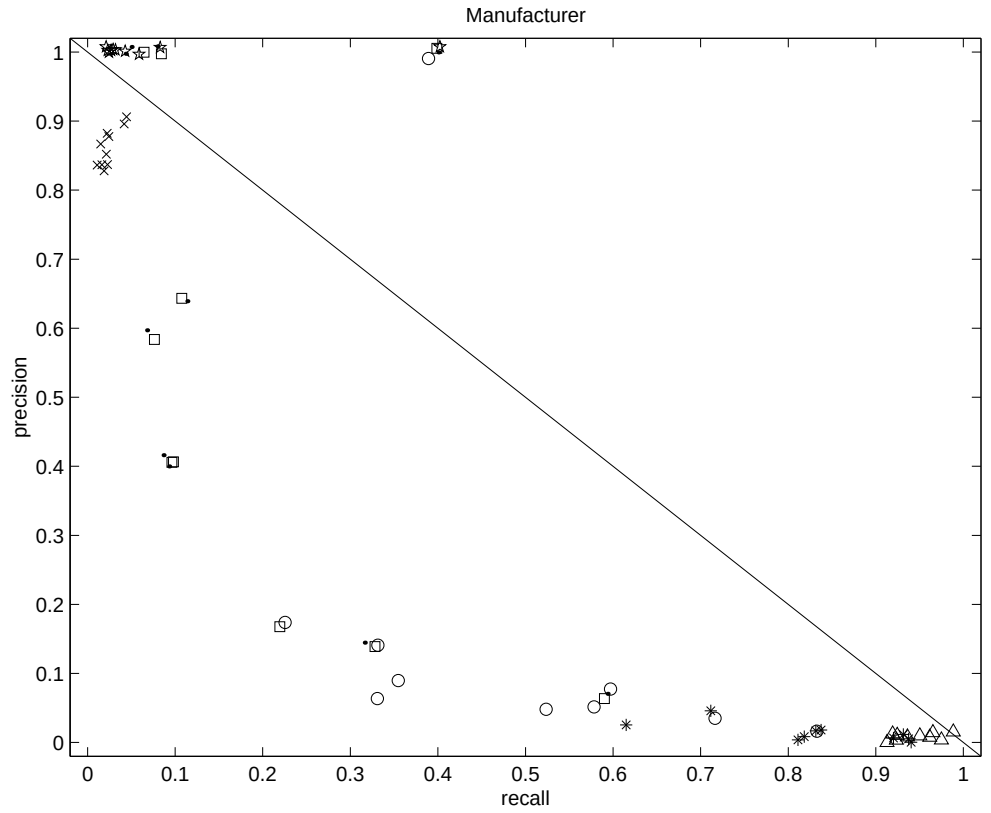

Mobile Phone
precision
recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

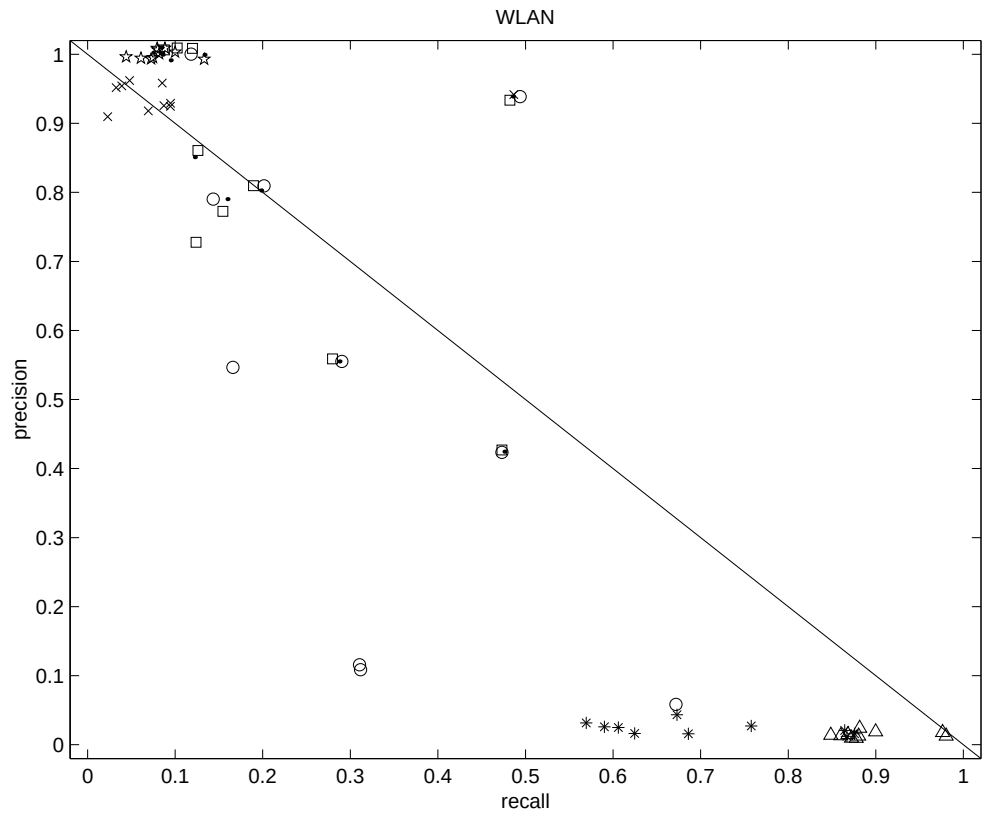

Investment/Investor
precision
recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

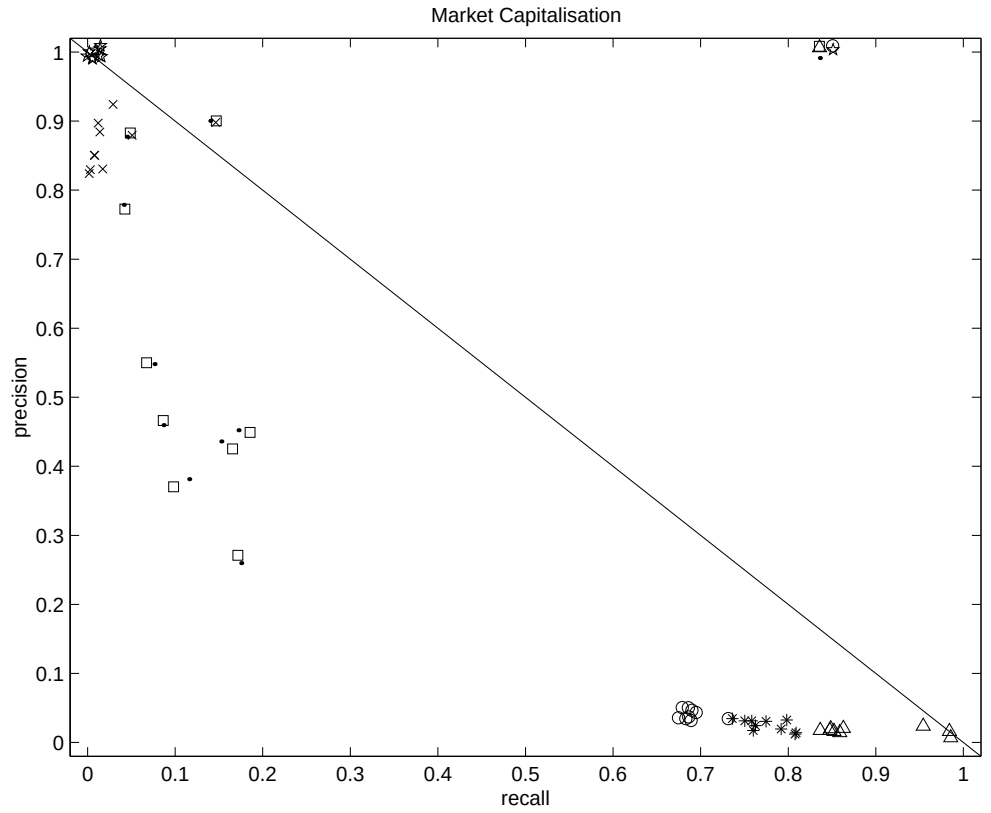

Network
precision
recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

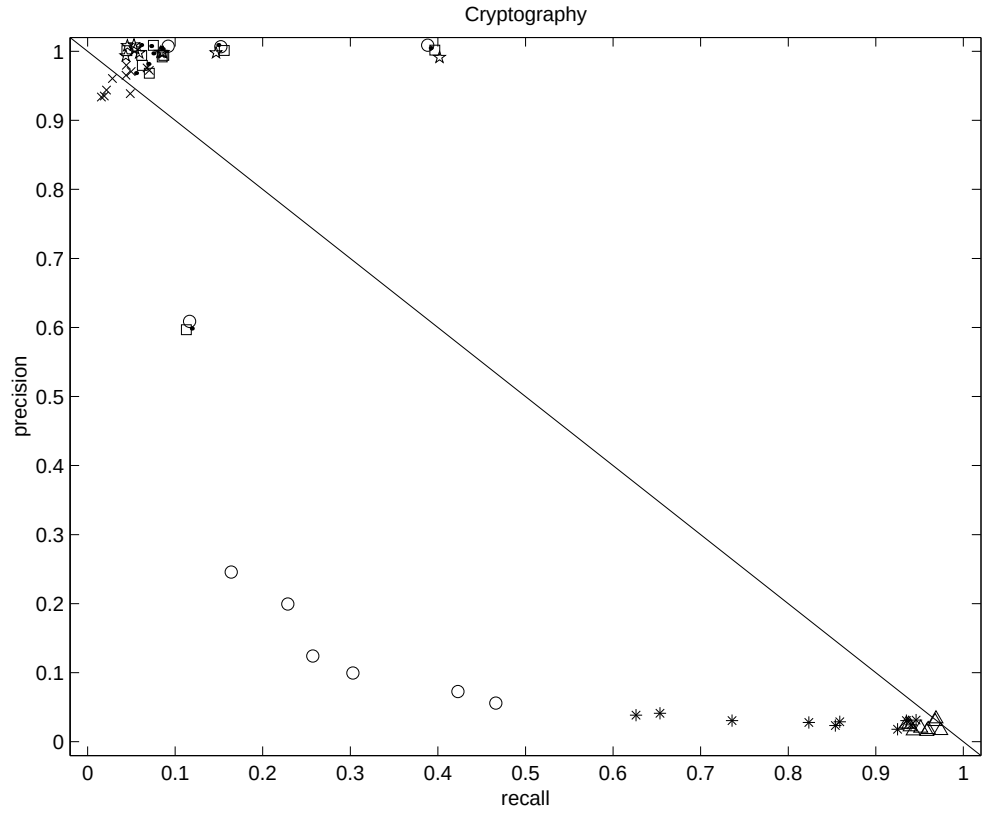

Artificial Intelligence
precision
recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

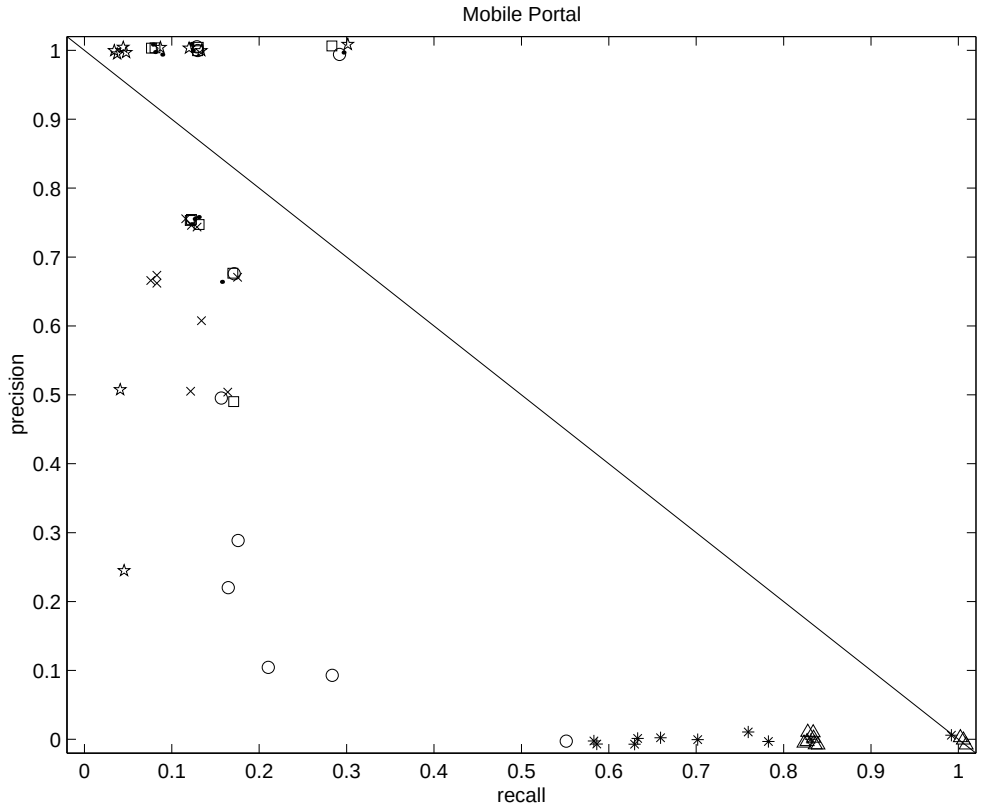

Online Community
precision
recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

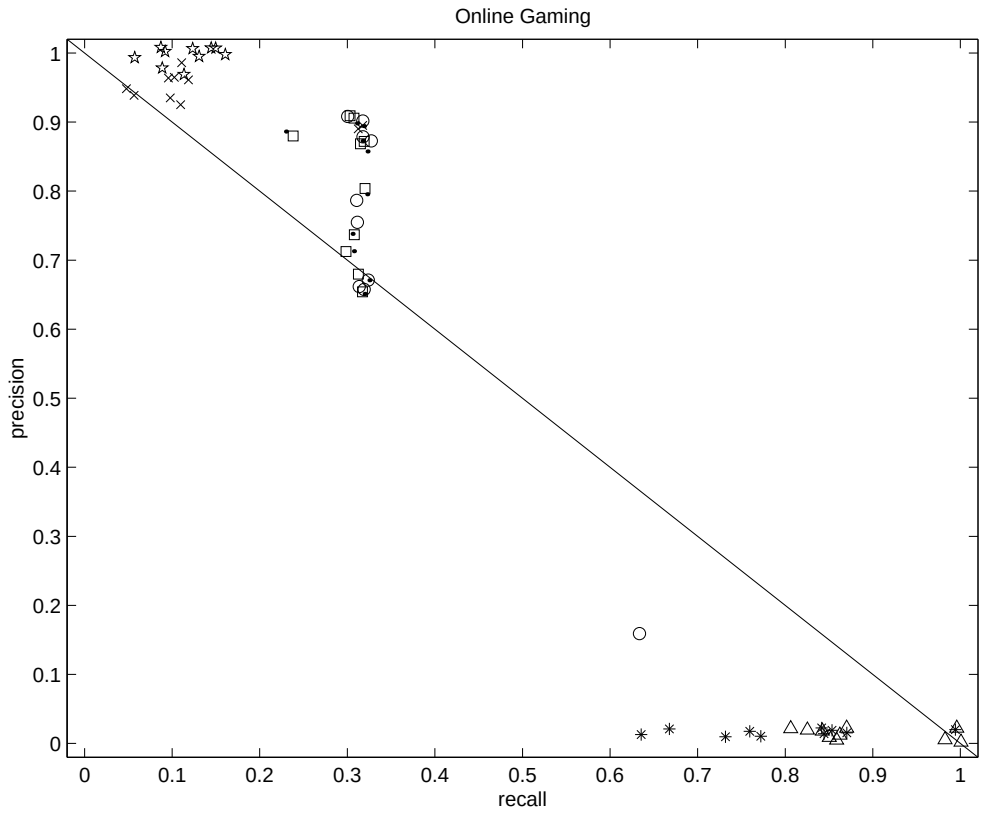

Manufacturer
precision
recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

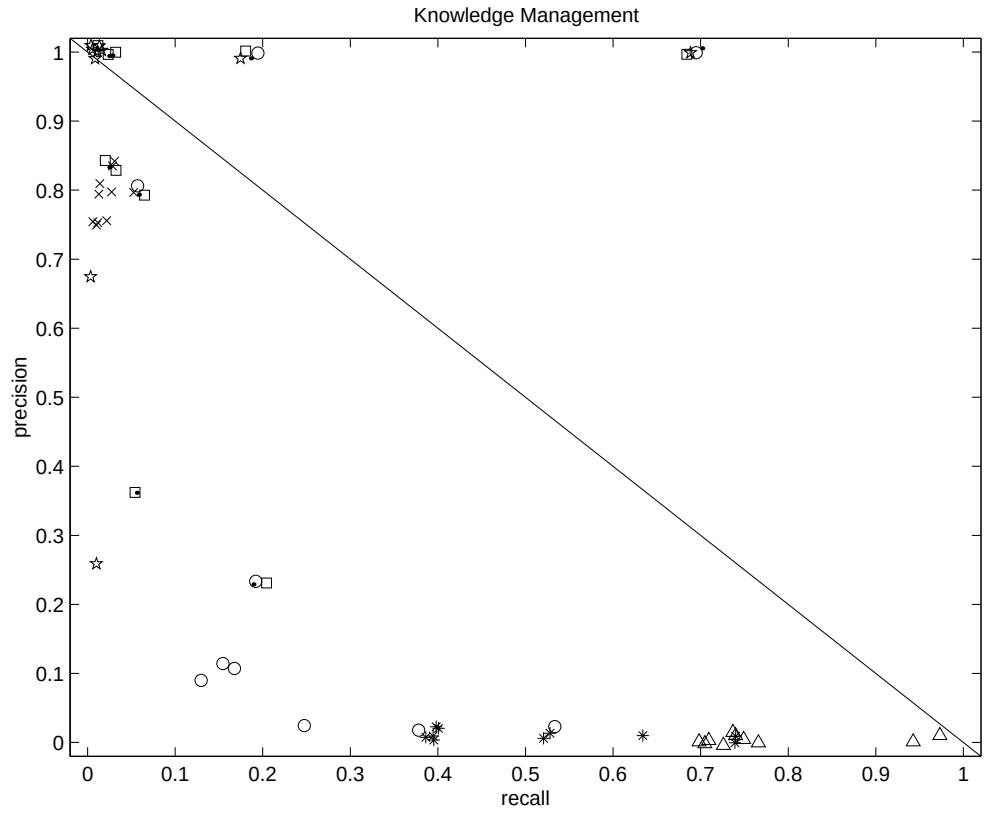

WLAN
precision
recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

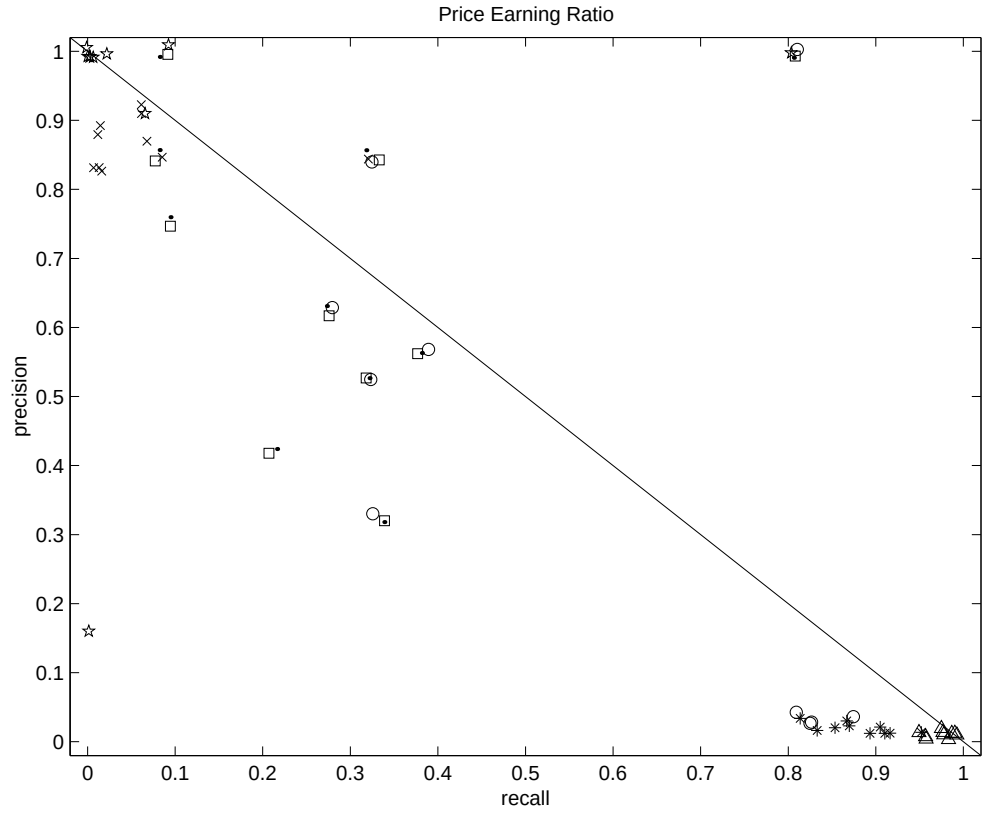

Market Capitalisation
precision
recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


Cryptography
precision
recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


Mobile Portal
precision
recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


Online Gaming
precision
recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


Knowledge Management
precision
recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


Price Earning Ratio
precision
recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1