Evaluation of Term Utility Functions for Very Short Multi-Document Summaries

Alexander K. Seewald, Christian Holzbaur

Austrian Research Institute for Artificial Intelligence,

Freyung 6/6, A-1010 Vienna, Austria

`{alexsee,christian}@oefai.at`

Gerhard Widmer

Department of Computational Perception,

Johannes Kepler University Linz

Altenberger Straße 69, A-4040 Linz, Austria

`gerhard.widmer@jku.at`

April 27, 2005

Abstract

We describe results from an application for relevance assessment in a setting related to multi-document summarization. For the task of characterizing given document collections by a short list of relevant terms, we have proposed the term utility function PxR. The measure is competitive to a variety of utility functions commonly used in text mining. Our function incorporates a user-definable parameter which allows for explicit, continuous trade-off between precision and recall, which was preferred by our users over the more opaque term utility functions from text mining. The F_β measure is similar but not identical to our measure and will also be discussed. Despite our users' preference for a user-definable parameter, the improvement by setting different user-defined parameter values for each document collection are limited, and a static value for the parameter works almost as well. This seems to be true for the F_β measure as well. A simple measure, SR, also performs competitively. In light of this evidence, a user-definable parameter seems to be unnecessary to achieve competitive performance.

1   Introduction

In this paper, we investigate the task of characterizing given document collections by a short list of relevant terms. This task is somewhat related to relevance assessment of topics in a multi-document summarization setting. However, our focus is on user interactivity, realtime feedback and understandability, and not on fully automatic approaches. In the context of our application we found a sim-

ple term utility function to be competitive to other common utility functions, while being preferred by the users of our system. Our measure shares some properties with the F_β measure, and we will discuss these similarities later.

For comparison, we considered a variety of common term utility functions from text mining, each of which maps every term to a numeric value which signifies the usefulness of the respective term to decide if documents are part of the respective collection or not. We compare the approaches twofold: via simple matching to the indexing patterns which were originally used to create the document collection and by referring to our users for manual evaluation.

We will first describe the application context, followed by our new measure and other common measures from IR. Then, we will give an overview about a set of document collections related to the topic of work which was provided by our partner, the Institute for Social Research and Analysis (SORA,`www.sora.at`). The computation of term statistics was done within the product Melvil by the Austrian company uma information technology AG (`www.uma.at`). These collections form the base for our experimental evaluation.

Afterwards, we will describe the experimental setup, discuss experimental results in the Results section, discuss earlier experiments and other issues in Discussion, give a short overview on related research, and finally conclude the paper.

2   Application Context

Within the EU IST project 3DSearch we investigated intelligent ways to improve Melvil, an ontology management tool by uma information technology AG. Detailed background on the application context as well as on other research within 3DSearch can be found in (Fürnkranz et al., 2002).

Within Melvil, an ontology is a hierarchical structure of connected concepts. Each concept corresponds to a collection of documents dealing with a specific topic, e.g. the *internet*, *wall street* or *artificial intelligence*. Each concept, or document collection, is described by a human-readable topic description and a regular expression, the latter of which is applied to a corpus of full-text documents downloaded from selected sources on the internet in order to retrieve documents which are concerned with the given topic.[1] Regular expressions take the form of multiple patterns, which are combined via logical OR, i.e. each pattern specializes on a subset of documents relevant for the collection. The union of the search results from all patterns yields the final collection. Patterns are themselves regular expressions and may contain subpatterns; however, this feature is seldom recorded. Examples for simple patterns are e.g. `\bmanpower\s+demand\b`, `\bsocial\s+security\s+contributions\b` and `\bSozialabgabe\w*`.

Creating ontologies is a time-consuming task which occupies a lot of the users' time. Quite a few iterations are necessary to achieve reasonably good

---

[1]The term ontology may be misleading, since the connections between concepts are arbitrary and all concepts' regular expressions are locally stored and independent of each other.

document collections – e.g. for ontology Arbeit 400 iterative changes were observed. Not all search terms are obvious choices, or even in the same language. Configuring additional internet sources for document retrieval may necessitate changes in many patterns.

So, in order to help users save time, we investigated iterative ontology improvement. The idea was to take a given document collection and characterize it by a short list of relevant terms. This list of relevant terms may suggest additional word patterns to the end-user, which are already implicitly present in the previously collected document collections.

Three additional issues were to be addressed: Real-time feedback (i.e. generation of lists of relevant terms), user interaction and comprehensibility of the measure and its parameters to the non-technical user. We believe that our proposed term utility function deals with all these constraints in an appropriate manner. It should be noted that user interaction does not seem to improve system performance significantly, and without user-interaction the other mentioned issues are no longer relevant.

User feedback from the Institute for Social Research and Analysis (SORA), based on relevant term lists by our proposed measure proved to be very positive, and detailed results will be reported later in this paper.

Table 1: This is the contingency table for term t and concept Co. a,b,c,d are the number of documents in the four categories along two independent dimensions: term occurrence and concept membership. t, contains term; ¬t, does not contain term; Co, is part of concept, ¬Co, is not part of concept.

	t	¬ t
Co	a	b
¬Co	c	d

3   Term Utility Functions

We propose a simple term utility function, PxR, based on explicit trade-off between precision and recall, where t stands for a term and Co for a given concept. Table 1 explains the variables a-d, and their relation to t and Co.[2]

$$PxR(t, Co) = Precision(t, Co)^x * Recall(t, Co)^{2-x} = \left(\frac{a}{a+c}\right)^x * \left(\frac{a}{a+b}\right)^{2-x}$$

(1)

For $x = 0$ the formula is equivalent to *recall*; for $x = 2$ it is equivalent to *precision*. In between the two extreme values, the function allows for continuous trade-off between precision and recall, chosen by the user. This has several advantages for our application:

- By efficiently pre-computing precision and recall for a given document collection and set of terms, we can instantly compute our measure for any

---

[2]Initially, we were inspired by a term utility function called PR, i.e. *precision* multiplied by *recall*, and generalized it to this term utility function. This also explains why we used $2 - x$ rather than $1 - x$, so that for $x = 1.0$ this function is exactly PR rather than its square root.

value of x. Thus, real-time feedback to the user becomes feasible.

- While the results from other term utility functions are sometimes hard to understand and explain, precision and recall are well-known concepts for many users, yielding a clear conceptual interpretation.

- Instead of coarse-grained user interaction (= choosing among a small set of known measures), our measure offers fine-grained user interaction (= choosing a continuous parameter x), where small changes in the parameter yield small changes in the resulting list of most relevant terms.

For comparison, we chose the following measures of term utility, most of which are commonly used for text-mining. All measures except one can be computed directly from the contingency table which is described in Table 1. Since the contingency table only captures term occurrence and not term frequency, we also calculated the sum of term frequencies for documents inside (f_c) and outside ($f_{\neg c}$) the given concept Co.

In initial experiments we found that $Precision$ $(\frac{a}{a+c})$ alone is unsuitable for term selection since many terms have the maximum precision of 1.0 ($a > 0$ and $c = 0$), which makes it impossible to determine a stable relative ranking, so we did not choose it for evaluation. However results for our measure with $x = 2.0$ – where ties are broken by preferring higher recall – address this problem, and yield a stable ranking of good performance for $Precision$ as well.

The following seven measures were considered for comparison. We have reformulated all measures as functions of values a-d from Table 1 and sometimes

also simplified the formula in a way which should not change the obtained ranking, e.g. removing outermost monotonic functions.

- χ^2 which determines whether there is a statistically significant relation between term occurrence and concept membership (Yang & Pedersen, 1997), i.e. $\frac{N(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$, where $N = a + b + c + d$ is the total number of documents.

- Information Gain (IG) which determines the information gained for concept prediction, given term occurrence; i.e. $-\frac{a+b}{N}log_2(\frac{a+b}{N}) + \frac{a}{N}log_2(\frac{a}{a+c}) + \frac{b}{N}log_2(\frac{b}{b+d})$. Both IG and χ^2 were found to be superior to all other considered features in (Yang & Pedersen, 1997).

- *oddsRatio*, which is a commonly used feature in information retrieval (Rijsbergen et al., 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. $\frac{a+c}{N}\log_2\frac{ac+ad}{ac+bc}$. It is equivalent to *FreqLogP* in (Mladenic, 1998).

- Recall (*recall*) is the ratio of documents which include the term, among all documents belonging to the concept, i.e. $\frac{a}{a+b}$.

- F-Measure (F_1), a static trade-off between recall and precision, i.e. $\frac{2*prec*recall}{prec+recall}$

- SimpleRatio (SR) is $\frac{f_c}{f_{\neg c}+1}$, which prefers those terms appearing frequently within the concept, but seldom without.

We are aware that a more general form of the F-Measure, F_β (for a derivation starting at F_1 see (Rennie, 2004)) is similar to our PxR in that it also has a user-definable parameter which controls trade-off between precision and recall. This will be discussed in a subsection of Results.

- F-Measure β (F_β), is a dynamic trade-off between recall and precision, i.e. for a given β between 0 and ∞, $F_\beta = \frac{(\beta+1)*prec*recall}{prec*\beta+recall}$. For $\beta = 1$, we get F_1 where *precision* and *recall* are equally weighted. F_0 is equivalent to *precision* while F_∞ is equivalent to *recall*. In some variants, β is squared in the given formula, i.e. $F_\beta = \frac{(\beta^2+1)*prec*recall}{prec*\beta^2+recall}$. This does not change the formula qualitatively, and amounts to changing the arbitrary exponential step of 10 which we use throughout this paper to an equally arbitrary step size of $\sqrt{10}$, and would have no effect on the main conclusions. Note especially that this new β amounts to the square root of the old β.

4 Experimental Setup

4.1 Ontology Arbeit

Our experimental evaluation is based on an ontology called Arbeit which was provided by the Institute for Social Research and Analysis.

The ontology contains 209 concepts of various complexity and sizes. Our users chose 10% (21) concepts for detailed analysis and later manual evaluation. Each concept is characterized by a set of patterns which have been initially created and iteratively refined by users over a period of several months,

Table 2: This table shows the 21 concepts which were chosen by our users for detailed evaluation. The columns show the (german) concept name, count of assigned unique documents, count of distinct patterns, and avg±stdDev of pattern length in characters.

Concept name	Docs	Patts	Avg.Len
OECD Raum	21,141	64	13.1±5.4
Technischer Wandel	8,702	51	22.3±5.6
Migranteneinrichtungen	6,115	20	24.8±12.3
Lohn/Einkommen	5,290	121	19.8±5.9
Interessensvertretungen/Kultur/Sport	5,009	18	16.3±7.5
Realitäten/Forschung	4,865	6	17.2±4.4
Sozial/Geisteswissenschaften	3,920	18	18.8±5.2
Österreichische Ministerien	3,522	37	45.6±24.5
Familienbeihilfe/Kindergeld	3,355	5	18.8±2.7
Weiterbildung und Qualifikation	3,128	29	21.7±7.6
Ältere Beschäftigte	2,985	36	24.3±9.3
ausländische Beschäftigte	2,154	58	20.5±8.0
Sonstige	906	4	9.8±1.3
Jugendeinrichtungen	735	13	18.7±8.3
Senioreneinrichtungen	598	6	22.5±6.4
Sonstige Beratungsinstitutionen	580	17	25.8±12.2
Arbeitskosten	575	19	21.3±7.3
Bildungsökonomie	495	24	24.1±6.0
Krankenversicherung	455	7	25.3±8.7
Eisen/Metall/Elektro	256	11	22.5±7.5
Niedrigqualifizierte	182	10	24.9±8.8

and several hundred steps of iterative refinements. Although these patterns are considerably more advanced than anything our term-based approach offers, they

are still a valuable resource for evaluating our approach. In Table 2 we show some details on the chosen concepts, i.e. the (mostly german) name, the count of assigned unique documents, the number of unique patterns and the average pattern length (in characters). The latter values are an indication of concept complexity, e.g. *Lohn/Einkommen* is quite complicated with 121 distinct patterns, which reflects the variety of possible income sources in Austria; while *Österreichische Ministerien* does not have so many distinct patterns, but easily the longest ones, since most Austrian ministries have very long names. A total of 69,396 unique documents were assigned to these twenty-one concepts. The overlap[3] over all these concepts is 1.08 and thus quite small – an indication that the concepts are well-defined and almost mutually exclusive. For comparison, within the research_march15 ontology from our earlier paper (Seewald et al., 2002), also mentioned in Section Discussion, the ten largest concepts had an overlap of 4.18, and the total overlap for all concepts was 20.08.

A total of 2,694,852 terms were present in the term index file for ontology Arbeit. This large number can be explained since Melvil uses *all* alphanumeric character sequences as terms for indexing, even if they only appear once. To reduce the vocabulary to manageable size, we removed all terms which appear in at most 10 documents, leaving us with 160,098 terms.

---

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

4.2 Evaluation Setup

A simplistic way to compare term utility functions would be to look at which highly-rated terms correspond to the indexing patterns which were used to define the concept. But since using single terms instead of regular expressions is a crude approximation at best, some information is inevitably lost, which leads to a systematic underestimation of true system performance. So, we considered two ways to compare our new measure:

- **Automatic Evaluation** We counted matches between the original indexing patterns (which were used to obtain the documents) and the top ten words selected by each measure. As we mentioned, the indexing patterns can be any regular expression. Thus, to allow for a fairer comparison, multiword indexing patterns were broken up into single word patterns at every place where a word boundary may appear, e.g. `metal\s+industry` maps to `metal | industry`.

- **Human Evaluation** (i.e. manual evaluation) We computed the top ten relevant terms selected by our PxR measure, for ten different values of x (from 0.0 to 1.8 in steps of 0.2 – 2.0 with tie breaking was added later due to reviewer feedback and not available for the original evaluation). The resulting list of 2,100 words was sent to our users. We asked them to count the number of relevant terms for each concept and value of x separately, and also decide on an optimal value for x, again separately for each concept. In some cases, a range of values were considered optimal and

Figure 1: This figure shows a scatterplot between p_c and p_h with fitted least-squares regression line ($r = 0.51$). Note the coarse-grained structure of both p_c and p_h which is caused by their definition – i.e. both can only attain multiples of 0.1.
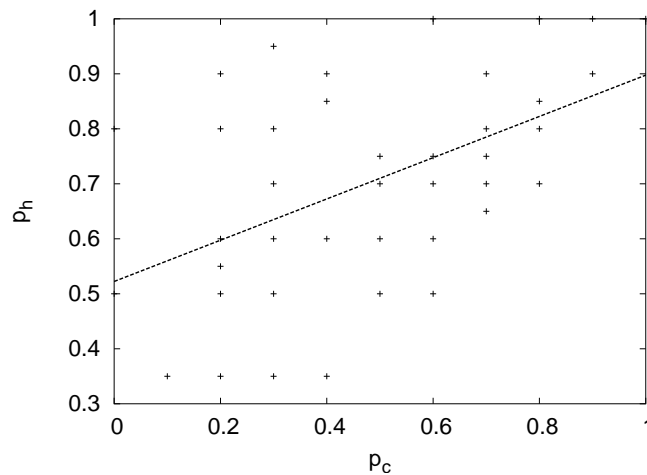
indistinguishable - in that case we took the arithmetic average of the minimum and maximum values within the range, rounding up as appropriate. As SORA is no longer available for further evaluation and additionally human evaluation is usually costly vs. an automatic evaluation, we have also investigated the relation between the results of automatic and human evaluation (see next section).

4.3   Human evaluation vs. automatic evaluation

For part of our data, we have both the automatically computed proportion of matched terms vs. indexing patterns, and the human judgement on the true

13

proportion of matching terms (i.e. those which are useful search terms for the given concept). Since human judgement is costly, and in our case SORA is no longer available for evaluation, we investigated the relation between both values under a simple setting. We call the computed proportion p_c and the human-judged true proportion p_h to facilitate this discussion.

First, the proportion of matched terms is always between 0.0 and 1.0 in steps of 0.1, as only ten terms are given for each concept and parameter value of our measure. One simple model for a relation between p_h and p_c is therefore a constant offset, i.e. $p_h = p_c + B$ where B is chosen as to minimize mean squared error (i.e. $\sum (p_h - p_c)^2$).

A more complex model would be to assume a linear relationship, i.e. $p_h = A * p_c + B$, where A and B is chosen as to minimize mean squared error. This model subsumes the first case when $A = 1$. Pearson's correlation coefficient is one way to measure the agreement of such a model, and allows to compute the regression line defined by A and B explicitly. Usually the scatterplot is inspected first to see whether a linear relationship is warranted, see Fig. 1.

Pooling judgements for all our twenty-one concepts, we get a total of 40 unique samples, each with a unique combination of p_c and p_h value. Fig. 1 shows the scatterplot of these computed term proportions (p_c) on the X axis and human-judged term proportions (p_h) on the Y axis plus the regression line $(A = 0.3752, B = 0.5224)$. One can see that p_c and p_h are only weakly correlated. Pearson's correlation coefficient of $r = 0.51$ $(r^2 = 0.26)$ shows that there is a slight linear relationship between p_h and p_c. A Fisher's t Test for

14

r confirms this relationship as significant at 5% confidence level. However, as $r^2 = 0.26$, only 26% of the variance is shared between p_c and p_h.

Least-squares linear regression on this data gives us a model to predict p_h from p_c ($A = 0.3752$, $B = 0.5224$). The square root of mean squared error (divided by the number of samples) for this model is 0.164 and mean absolute error is 0.132. So we must expect an average error in p_h of about 0.1-0.2, or 1-2 terms. As $A = 0.3752$, this translates back into an expected average error in p_c of 0.27-0.53 or 3-5 terms. I.e. we expect that changes in p_c which are smaller than 0.53 do not influence the estimated p_h beyond the average error of the linear model. Only differences of more than 5 terms can thus be considered as significant beyond linear model's uncertainty. This is not precise enough to distinguish any two of our measures.

Computing the correlation for each concept separately gives only a single significant relation for concept *Senioreneinrichtungen* with $r = 1.0$. All others are either not significant at 1% confidence level, or do not have more than two unique samples[4]. Reducing the confidence level to 5% still gives only one significant per-concept relation. Adding to that, *Senioreneinrichtungen* has only three samples – exactly the minimum size necessary – so we are inclined to see this result as a statistical fluke. In any case one concept would be too little data for fitting local linear models. Local linear models may have worked better but would have made the analysis susceptible to overfitting due to the much higher

---

[4]It is always possible to run a regression line through zero, one or two points with perfect correlation.

number of parameters to be fitted from our limited data.

Concluding, we have found that there is a slight linear relationship between p_h and p_c, so there is some merit in using p_c to replace the costly human evaluation to obtain p_h. However, the correlation is not strong and errors are high: only a difference of more than 0.5 in p_c can be considered significant beyond linear modelling uncertainty. High performance according to automatic evaluation via p_c is therefore not necessarily a sign for high performance according to human evaluation via p_h, since only about a quarter of the variance is shared between these variables.

5 Results

5.1 Human evaluation

As mentioned previously, SORA evaluated our model on the twenty-one concepts with the top ten terms from each $x = 0$ to $x = 1.8$ in steps of 0.2. Note that SORA did not receive the $x = 2.0$ model which was only implemented recently due to reviewer feedback. They defined relevance as any term which might prove useful to add as search term to the given concept, including those terms which were already present as partial or full indexing patterns.

They found that our measure averages 8.1 ± 1.6 relevant terms in the top ten. The approach to count matches automatically by computing overlap with indexing patterns underestimates the performance as expected: The automatic approach estimates 5.9 ± 2.7 matched terms for the optimal values of x deter-

16

mined by the users[5]. The best overall setting of $x = 1.8$ gives 7.5 ± 2.0 relevant terms in the top ten according to human evaluation, and 5.9 ± 3.0 according to automatic evaluation.

Contrary to our expectations, not much could be gained by adapting the parameter x to each concept separately: 8.1 vs. 7.5 matches in the top ten terms, which means roughly half a significant term more – not much indeed when you consider that this means looking at roughly an order of magnitude more terms. A fixed parameter value works almost as well, which may indicate that simple measures such as SR (with similar performance in the automatic evaluation, see next section) may work as well as our measure. Unfortunately, SORA is no longer available for evaluation so we cannot check this thoroughly.

5.2 Automatic evaluation

Figs. 2 and 3 show the averaged results as proportion of the top ten terms selected by each measure which matches any of the indexing patterns. Matching is done automatically and is case-insensitive. Both figures also show the combined results, when choosing the optimal measure resp. parameter value for each concept separately, by the benefit of hindsight. For comparison, Fig. 3 also shows the performance of our system as evaluated by the users (at the far right). This evaluation is the only one in this figure which is *not* based on term matches with indexing patterns. Complete details can be found in Tables 3 and

---

[5]which were 1.6, 1.0, 1.8, 1.8, 1.8, 1.2, 1.8, 1.6, 1.8, 1.8, 1.6, 1.6, 1.4, 1.2, 1.6, 0.8, 1.8, 1.6, 1.4, 0.8, and 1.2 (1.49 ± 0.33) in the order of Table 2

Figure 2: This plot shows the average fraction of top ten terms selected by each measure which match any of the indexing patterns. The common IR measures χ^2, *IG*, *oddsR*, *odds2*, *recall*, F_1, and *SR* are shown. *best* to the right of the dotted line shows the combined results when choosing for each concept the optimal measure by hindsight.

Common IR Measures
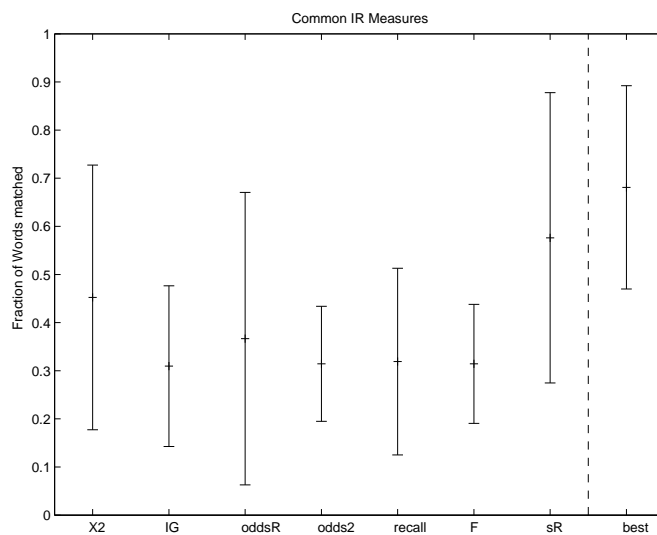
4.

We see that *PxR* is competitive to earlier approaches. Generally, *SR* and χ^2 seem to be the best IR measures, and *PxR* with $x >= 1.2$ yields similar performance, peaking at around $x = 1.8$ with almost the same average performance as *SR*. The simple measure *SR* thus performs surprisingly well.

Figure 3: This plot shows the average fraction of top ten terms selected for each value of the parameter x (except 2.0 which was not available to SORA) by PxR which match any of the indexing patterns. The leftmost entry in *best* to the right of the dotted line shows the combined results when choosing for each concept the optimal value of x by hindsight. The rightmost entry in *best* shows our users' evaluation of the optimal fraction for each concept and is the only one *not* based on comparison with the indexing patterns.
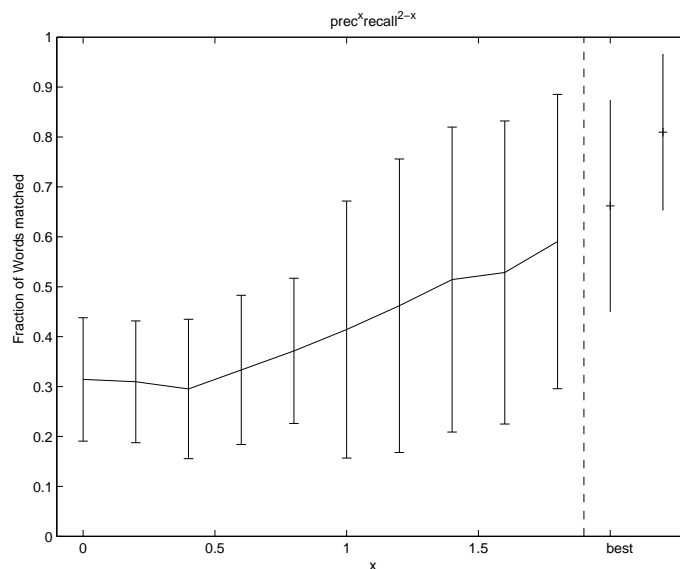
5.3 Comparing to F_β

The F_β measure (Rennie, 2004) is similar to our measure in that it also has a parameter which can be interpreted as trade-off between precision and recall ($\beta < 1$ gives *precision* more weight and $\beta > 1$ gives *recall* more weight).

A disadvantage of F_β is that the interval for parameter β is not bounded

19

Table 3: This table shows the computed proportion of matched terms vs. indexing patterns for common IR measures. Average and standard deviation over all concepts are also given.

Concept name	χ^2	IG	oddsR	odds2	F	recall	SR
OECD	0.8	0.7	0.8	0.7	0.1	0.1	1.0
TW	0.2	0.3	0.1	0.4	0.2	0.2	0.6
ME	0.4	0.5	0.2	0.2	0.3	0.1	0.8
L/E	0.1	0.3	0.6	0.3	0.0	0.4	0.1
Iv/K/S	0.6	0.5	0.5	0.3	0.4	0.3	0.9
R/F	0.4	0.5	0.6	0.4	0.4	0.5	0.7
S/G	0.2	0.2	0.3	0.1	0.2	0.1	0.7
ÖM	0.3	0.2	0.3	0.3	0.1	0.4	0.6
FBH/KG	0.5	0.3	0.5	0.4	0.3	0.2	0.6
W&Q	0.6	0.4	0.9	0.3	0.4	0.4	0.8
ÄB	0.7	0.2	0.7	0.1	0.4	0.2	1.0
AB	1.0	0.3	1.0	0.4	0.6	0.2	1.0
S	0.3	0.3	0.2	0.3	0.6	0.3	0.2
JE	0.3	0.2	0.1	0.3	0.3	0.4	0.2
SE	0.9	0.3	0.3	0.3	0.7	0.4	0.7
SB	0.0	0.2	0.0	0.3	0.0	0.4	0.0
AK	0.2	0.0	0.0	0.3	0.3	0.4	0.2
BÖ	0.5	0.0	0.1	0.3	0.3	0.4	0.5
KV	0.2	0.3	0.0	0.3	0.2	0.4	0.4
E/M/E	0.5	0.5	0.2	0.3	0.3	0.4	0.5
NQ	0.8	0.3	0.3	0.3	0.6	0.4	0.6
Avg.	0.45	0.31	0.37	0.31	0.32	0.31	0.58
±stD	0.27	0.17	0.30	0.12	0.19	0.12	0.30

Table 4: This table shows the proportion of matched terms vs. indexing patterns for our *PxR* measure. The columns correspond to different values for the parameter x. For $x = 2.0$, ties were broken by preferring terms with higher recall. Average and standard deviation over all concepts are also given.

Concept	0.0	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6	1.8	2.0
OECD	0.1	0.1	0.1	0.1	0.2	0.1	0.4	0.9	0.9	0.9	1.0
TW	0.2	0.2	0.3	0.3	0.4	0.3	0.2	0.1	0.1	0.2	0.2
ME	0.1	0.2	0.2	0.2	0.3	0.6	0.5	0.5	0.5	0.7	0.8
L/E	0.4	0.4	0.3	0.3	0.2	0.0	0.1	0.3	0.3	0.3	0.3
Iv/K/S	0.3	0.3	0.4	0.5	0.6	0.6	0.6	0.7	0.7	0.8	0.9
R/F	0.5	0.5	0.5	0.5	0.5	0.6	0.3	0.3	0.4	0.7	0.8
S/G	0.1	0.1	0.1	0.4	0.2	0.2	0.2	0.2	0.3	0.6	0.6
ÖM	0.4	0.4	0.4	0.3	0.3	0.3	0.3	0.4	0.5	0.6	0.5
F/K	0.2	0.3	0.3	0.3	0.4	0.4	0.6	0.6	0.6	0.6	0.3
W&Q	0.4	0.4	0.3	0.4	0.5	0.6	0.6	0.6	0.7	0.8	0.8
ÄB	0.2	0.1	0.2	0.3	0.3	0.5	0.9	1.0	1.0	1.0	1.0
AB	0.2	0.3	0.3	0.3	0.5	0.8	1.0	1.0	1.0	1.0	1.0
S	0.3	0.3	0.5	0.4	0.4	0.3	0.2	0.2	0.2	0.2	0.0
JE	0.4	0.4	0.3	0.3	0.2	0.3	0.3	0.3	0.2	0.2	0.1
SE	0.4	0.4	0.4	0.5	0.6	0.9	1.0	1.0	1.0	1.0	0.8
SB	0.4	0.4	0.2	0.3	0.2	0.0	0.0	0.0	0.0	0.0	0.0
AK	0.4	0.3	0.2	0.1	0.3	0.2	0.3	0.3	0.3	0.4	0.4
BÖ	0.4	0.2	0.1	0.0	0.4	0.5	0.8	0.8	0.8	0.8	0.8
KV	0.4	0.5	0.5	0.4	0.2	0.2	0.2	0.4	0.4	0.4	0.4
E/M/E	0.4	0.4	0.5	0.6	0.5	0.5	0.5	0.5	0.5	0.5	0.3
NQ	0.4	0.3	0.1	0.5	0.6	0.8	0.7	0.7	0.7	0.7	0.4
Avg.	0.31	0.31	0.30	0.33	0.37	0.41	0.46	0.51	0.53	0.59	0.54
±stD	0.12	0.12	0.14	0.15	0.15	0.26	0.30	0.31	0.30	0.30	0.33

Figure 4: This is a comparison of PxR (on the left) and the F_β measure (on the right) as a function of *precision* (X axis) and *recall* (Y axis). $x = 0/0.5/1/1.5/2$ is shown as well as $F_{100/10/1/0.1/0.01}$. Isolines correspond to values of each function of 0 to 1 in steps of 0.05. Isolines for 0 and 1 may be aligned with the axes and therefore invisible. Sharp bends in some isolines are artefacts of gnuplot's linear sampling.
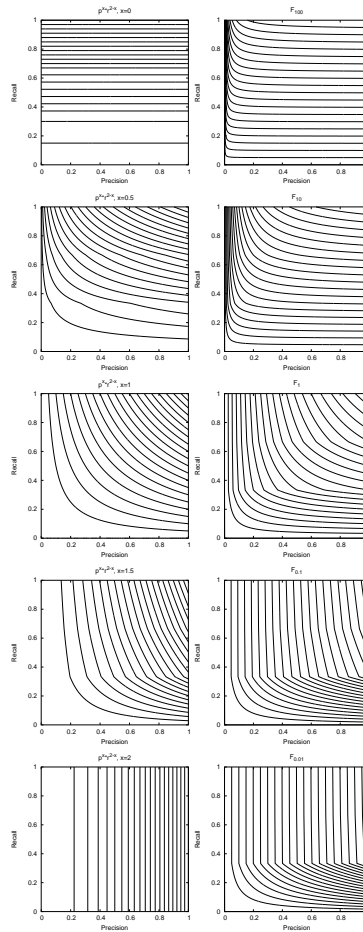
Figure 5: This plot shows the average fraction of top ten terms selected by the measure F_β. β values from 10^{-5} to 10^5 were tested in steps of 10 on an exponential scale. Note that due to the different interpretation of the β parameter small values of β prefer precision over recall just like large values of x for PxR and vice versa, so that the order on the X axis has the opposite meaning versus Fig.3.
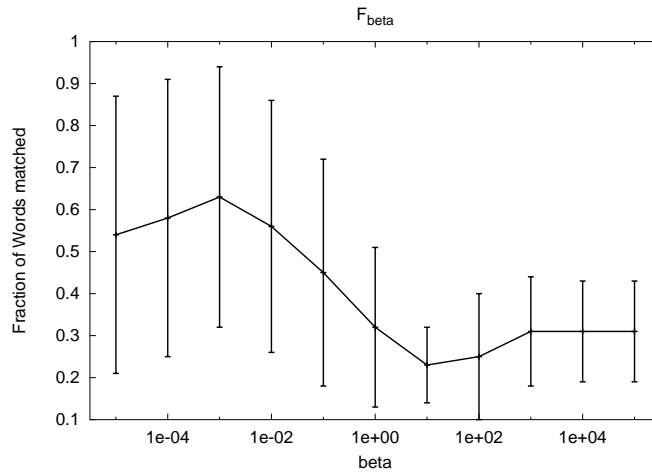
on one side and may become arbitrary large, while for PxR the range is contained in the closed interval $[0, 2]$. Closed intervals are easier to visualize in a user interface, and may be better suited for people without mathematical background.

Fig. 4 shows a side-by-side comparison of PxR and the F_β measure for various values of x and β. A β value of zero corresponds to $x = 2$ in that only precision determines the output while a β value of ∞ corresponds to $x = 0$ in that only recall determines the output. $\beta = x = 1$ corresponds to equal weight for

23

precision and recall in both measures (third row in the figure). The form of the isolines from both measures is similar but not identical. Instead of $\beta = 0$ we used $\beta = 0.01$ and instead of $\beta = \infty$ we used $\beta = 100$. This was necessary as we wanted to keep the plots symmetric between top and bottom, although the geometric step between adjacent βs is arbitrary.[6]

Fig. 5 shows the average proportion of recovered indexing terms for F_β, similar to Fig. 3 which shows the same property for our PxR measure. Full results are found in Table 5. As can be seen, both measures perform comparable, and F_β even performs slightly better than the best value of PxR at $\beta = 10^{-3}$. This does not necessarily ensure a good performance according to human evaluation since human and automatic evaluation are only weakly correlated.

Although the F_β measure and PxR seem remarkably similar, there is in general no way to compute a β from a given x so that the same ranking appears for both measures, except for the extremal points we discussed earlier. Even for $\beta = x = 1$, the ranking would still be different in almost all cases. The measures are similar, but *not* identical.

6   Discussion

We will now discuss the relation of our experiments here with earlier experiments reported in (Seewald et al., 2002). In the mentioned paper, we used a different ontology, the research_march15 ontology. This ontology was built by

---

[6]With a factor of 100, i.e. $\beta = 0.001$ and $\beta = 10000$, the plots looked very similar to those shown here.

Figure 6: Each point corresponds to a term for the measures $\bullet = \chi^2$, $\circ = IG$, $\times = oddsRatio$, $* = odds2$, $\square = PR$, $\triangle = recall$ and $\star = sR3$. The precision and recall of each term determines its position within the graph. Only the top ten terms of each measure are shown. Absolute noise (jitter, ± 0.01) was added to improve visualization.
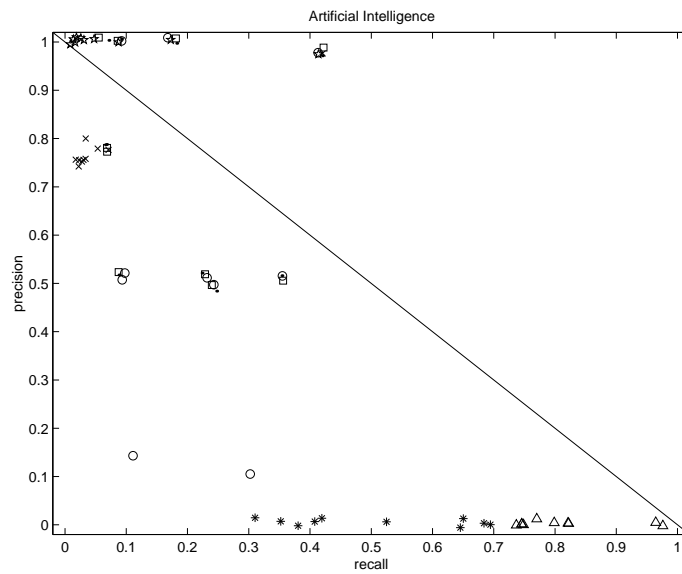
the technicians of Melvil for demonstration purposes. We choose the ten largest concepts, and arbitrarily ten smaller concepts, from the ontology. No evaluation via PxR took place, but we compared mainly the same set of measures from information retrieval which were used here.

Furthermore, we were restricted to a purely automatic evaluation similar to the one explained here. This earlier evaluation yielded much worse results: results for matched term proportions (p_c) ranged from 0.04 to 0.29 for the standard measures in information retrieval. As can be seen, the results presented

here are much better at 0.31-0.58. Also, a surprising result is that *SR* (called

sR3 in (Seewald et al., 2002)) performs **much** better here at around 0.59 versus

0.04/0.2 for the large/small concepts of research_march15. Measure *IG* which is

one of the worst-performing measure here performed at best (smaller concepts)

and second-best (large concepts) place on research_march15.

We have the following hypotheses to explain these discrepancies. We believe

the first hypothesis is the most significant one, although the other two may also

contribute, albeit to a smaller extent.

- The research_march15 ontology was not created to fulfil a specific purpose

  other than demonstrating the system. Therefore, the described concepts

  may have been less cohesive and consistent than in ontology Arbeit. This

  is supported by the high overlap of 4.18 for the largest ten concepts (20.08

  over all concepts) versus 1.08 for the 21 chosen concepts from ontology

  Arbeit.

- We have found that quite many terms which were indexed initially may

  be explained as internal html-tags[7] or other suspicious words[8] which are

  usually not considered part of human-readable text. In the time course

  these bugs may have been corrected.

- We also noticed earlier that the full text index of Melvil 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*. We presumed that the terms were

---

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

[8] e.g. 00000000000000000000000001, aaaaaa, abcdefghijklmnopqrstuvwxyz, etc.

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. This bug may have been fixed as well.

Contrary to this earlier study, where we mentioned that one cannot unequivocally say which of the measures is best, the choice here is obvious: either SR, or PxR with a default value of $x = 1.8$, followed with some performance deterioration by χ^2. F_β may also be an option, but has not been validated by human evaluation as PxR has. Only about five of the twenty-one concepts offer a better performance for at least one other measure. As we already mentioned, adapting the value of x for each concept does not improve performance by much and increases the workload disproportionately. We believe that the ontology Arbeit used here is more representative and the results presented here should hold more generally.

Finally, since this is the journal for Applied Artificial Intelligence, we would like to share the visualization of the concept *Artificial Intelligence* from the research_march15 ontology, see Fig. 6. The top ten terms for this concept from measure SR are ki, ai, trappl, dunietz, hutchens, verbmobil, seminarvorträge (short lectures), goren, treister and dfki; the terms from PR (PxR for $x = 1.0$) are ai, intelligenz, ki, artificial, künstliche (German for artificial), seminarvorträge, kampfroboter (robot fighters), wunderwaffe (wonder weapon),

arbeitssklaven (working slaves) and privatstiftungen (foundations). The idiosyncratic nature of these terms may be explained by an interesting view on Artificial Intelligence by Melvil's technicians, or by the small number of documents in this concept – only 239.

7   Related Research

The task we have addressed here is similar to both Task 1 (very short single-document summaries) and Task 2 (short multi-document summaries) from recent *Document Understanding Conferences* (DUC 2004, Workshop on Document Understanding, (Over & Yen, 2004)). What we have aimed for is related to a very short multi-doc summary (i.e. ten words to partially describe a multi-document collection). Our evaluation is biased towards the project's requirements and does not focus on creating a full summary of the document collection. However, it is interesting to note that up to date none of the current systems achieve better than baseline performance according to (Over & Yen, 2004). Baseline performance was provided by taking the first 75 bytes of each text for Task 1, and the first 665 bytes for Task 2. This confirms one common heuristic in summarization research, namely that important details are usually at the beginning of the first paragraph.

Our work here is more loosely related to the problem of keyphrase extraction, which is an important means for document summarization, clustering, and topic search. (Frank et al., 1999) gives some background on keyphrase ex-

traction and describes the open-source system KEA. They report that roughly one of five, or two of fifteen keyphrases returned by the system were equivalent to manually assigned keyphrases. The main differences to our system is that we only consider single word phrases while keyphrase extraction usually deals with multi-word phrases; and that we are dealing with multi-document summarization while keyphrase extraction deals with the summarization of a single document. Additionally our focus is not on completely characterizing a document collection, but to find related terms which may be used to extend it. This is obviously a simpler task, so it is not unexpected our results are much better: For *PxR*, around 7-8 of 10 terms proposed by our system were found to be relevant in a manual evaluation by domain experts.

(Jones & Paynter, 2003) discuss a manual evaluation of full sets of keyphrases by human domain experts, and is thus related to our human evaluation approach. It discusses the problem of automated evaluation using precision and recall, and gives an overview of earlier approaches to judge keyphrases one-by-one. They found that keyphrase sets by authors are ranked highest, followed by other human-generated set. Both have almost indistinguishable average scores (6.65 and 6.63 of 10). The third-best ranked set overall was a variant of the KEA system, with a 4.4% smaller score (6.20). This seems to indicate that properly trained keyphrase extraction systems come close to human performance. This is also what we found for our much simpler system for multi-document summarization, although we did not explicitly test human performance on the same task. Their human evaluation was more comprehensive in that they also con-

sidered coherence, discourse structure and consistency of the multi-document summaries, most of which cannot be applied to lists of relevant terms that we use.

Google News, `news.google.com`, is a current approach loosely related to multi-document summarization in the newspaper domain. However, the scientific methods underlying their approach are largely unknown, except that they involve clustering algorithms to group similar news articles. Therefore the summarization is more like a clustering of news documents which describe the same news message (with a choice on which one to take as representative and display), without any explicit summarization of message contents. This approach is probably helped by the fact that the set of news agencies is much smaller than the set of newspapers, and that articles by news agencies are often copied verbatim by many newspapers with only small changes in style and layout. It also indicates that the expected benefits from summarization may be limited. Combined with the fact that very large numbers of news sources and articles are continuously analysed, the clustering algorithm must be very efficient. Similarily, our approach works very efficiently by relying on simple counting of words within and without documents without complex lexical or statistical preprocessing. However, our approach is more strongly related to multi-document summarization than to clustering similar documents.

(Marcu, 2003) gives an excellent overview on past and present techniques for document summarization. He notes that current approaches in headline summarization, which are most similar to our task, use mainly statistical ap-

proaches from machine learning to estimate probabilities for given words being in the headline as well as to distinguish grammatical from ungrammatical headlines. Our approach differs in that only the contingency table for each word vs. the concept is computed and no learning takes place, and we also do not aim to create continous headlines. He also notes that static summarization systems differ radically in performance on different tasks, which we have also experienced – compare results from ontology research_march15 (section Discussion) vs. ontology Arbeit. He concludes that *Building a summarization system that is better than a dumb one that selects the first n sentences in a news article is still a significant challenge.*

(Marcu & Gerber, 2001) gives an overview on multi-document summarization, which is related to our task. They conclude that current evaluation protocols and techniques are not yet able to distinguish between good and bad systems. So, in a way our primitive approach to multi-document summarization as a short list of relevant terms may prove promising.

(Mani, 2001) provides an overview of different methods for evaluating automatic summarization systems. Our approach seems to be most similar to relevance assessment, where the relevance of terms for given topics (=concepts) is judged manually by human domain experts and also by automatic matching of terms to the originally used indexing patterns.

(Hovy & Lin, 1999) describe the SUMMARIST system for text summarization. The system is based on techniques from Information Retrieval and Extraction of single documents, which are extended with symbolic, semantic

and statistical methods. They have found that major topics within a document are clearly related to position of relevant words. Our work differs in that our terms are only very simple pseudo-topics, and that we do not use positional information to determine relevance.

(Mladenic, 1998) describes several known and some new methods for feature subset selection on large text data. They find that measures derived from odds ratio work best on a small dataset with text data from web page hyperlinks. For our task, odds ratio and their proposed derived measure both perform rather bad.

(Yang & Pedersen, 1997) offers a comprehensive comparative study on feature selection in text categorization. While they focus on dimensionality reduction for text categorization learning, we focus on user-interaction and real-time feedback. They found IG and χ^2 to perform superior to all other measures considered. We agree on χ^2 which yields a good second-best – however, IG performs very badly.

8 Conclusions

We have investigated a simple term utility function, PxR, based on user-defined trade-off between recall and precision. For our application fine-grained user-interactivity and the feasibility for real-time feedback was more important than performance. We have shown that our new function is competitive in an automatic evaluation to seven other measures commonly used in information re-

trieval tasks, given appropriate settings for parameter x. We have also shown that our new function performs well in human evaluation, returning on average 8 of 10 terms which are related to the ontology's concept. A comparison with the F_β measure showed that it is competitive to our measure and even performs slightly better in the automatic evaluation. Contrary to our expectations, a fixed value of $x = 1.8$ performed almost as well (7.5 vs. 8.1 matches out of ten) so user interactivity fails to strongly improve performance. In this light, simple non-interactive measures such as SR might be expected to be sufficient as well.

Although similar, the F_β measure was found to be non-identical except at two extremal points ($\beta = 0$, $x = 2$ and $\beta = \infty$, $x = 0$). Analyzing the results of both measures indicates that neither *precision* alone – even with tie breaking by preferring higher recall (PxR, $x = 2.0$) – nor the smallest weight we considered for *recall* (F_β, $\beta = 10^-5$) is sufficient to achieve best performance. Rather, an appropriate combination of *recall* and *precision* is superior, albeit with the much stronger weight on *precision*. A simpler alternative without user-definable parameter might be the SR measure which is not directly related to precision or recall, is easy and fast to compute and performs competitively.

Acknowledgements

References

Frank E., Paynter G.W., Witten I.H., Gutwin C. and Nevill-Manning C.G. (1999) Domain-specific keyphrase extraction. Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, Morgan Kaufmann Publishers, San Francisco, CA, pp. 668-673.

Fürnkranz J., Holzbaur C., Temel R.: User Profiling for the Melvil Knowledge Retrieval System. Applied Artificial Intelligence 16(4):243-281, April 2002.

Hovy E., Lin C.-Y. (1999) Automated Text Summarization in SUMMARIST. Advances in Automatic Text Summarization, I. Mani and M. Maybury (editors), 1999.

Jones S., Paynter G.W. (2003): An Evaluation of Document Keyphrase Sets. Journal of Digital Information, Volume 4 Issue 1. Article No. 122, 2003-02-19. `http://jodi.tamu.edu/Articles/v04/i01/Jones/`

Mani, I. (2001) Summarization Evaluation: An Overview. Invited Speaker to the

Automatic Summarization Workshop, Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL 2001)

Marcu, D. (2003). Automatic Abstracting, Encyclopedia of Library and Information Science, pp.245-256, 2003.

Marcu, D., Gerber, L. (2001) An Inquiry into the Nature of Multidocument Abstracts, Extracts and Their Evaluation, Proceedings of the NAACL-01 Workshop on Text Summarization, Pittsburgh, PA, 2001.

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

Over P., Yen J. (2004): An Introduction to DUC 2004 Intrinsic Evaluation of Generic New Text Summarization Systems. Proceedings of the Workshop on Document Understanding (DUC 2004), `http://duc.nist.gov/pubs.html#2004`.

Rennie, J.D.M. (2004) Derivation of the F-Measure, Feburary 2004, `http://people.csail.mit.edu/~jrennie/writing`.

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.

Seewald A.K., Holzbaur C., Widmer G.: Offline Evaluation of Term Utility Functions. Technical Report, Austrian Research Institute for Artificial Intelligence, Wien, TR-2002-34, 2002. `www.oefai.at`

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.

Table 5: This table shows the proportion of matched terms vs. indexing patterns for the F_β measure. The columns correspond to different values for the parameter β on an exponential scale. Average and standard deviation over all concepts are also given. To facilitate comparison with PxR, the values of β begin with the largest value and successively get smaller.

Concept	10^5	10^4	10^3	10^2	10	1	0.1	10^{-2}	10^{-3}	10^{-4}	10^{-5}
OECD	0.1	0.1	0.1	0.1	0.1	0.1	0.8	0.9	0.9	0.8	1.0
TW	0.2	0.2	0.2	0.2	0.2	0.2	0.1	0.2	0.3	0.2	0.2
ME	0.1	0.1	0.1	0.1	0.2	0.3	0.4	0.5	0.8	0.8	0.8
L/E	0.4	0.4	0.4	0.4	0.3	0.0	0.3	0.3	0.3	0.3	0.3
Iv/K/S	0.3	0.3	0.3	0.3	0.4	0.4	0.6	0.7	0.9	0.9	0.9
R/F	0.5	0.5	0.5	0.5	0.4	0.4	0.3	0.5	0.8	0.8	0.8
S/G	0.1	0.1	0.1	0.1	0.2	0.2	0.2	0.6	0.9	0.7	0.6
ÖM	0.4	0.4	0.4	0.4	0.3	0.1	0.4	0.5	0.7	0.6	0.5
FBH/KG	0.2	0.2	0.2	0.2	0.3	0.3	0.5	0.7	0.6	0.7	0.3
W&Q	0.4	0.4	0.4	0.4	0.3	0.4	0.6	0.8	0.8	0.8	0.8
ÄB	0.2	0.2	0.2	0.1	0.2	0.4	0.8	1.0	1.0	1.0	1.0
AB	0.2	0.2	0.2	0.3	0.2	0.6	0.9	1.0	1.0	1.0	1.0
S	0.3	0.3	0.3	0.3	0.3	0.6	0.2	0.2	0.2	0.0	0.0
JE	0.4	0.4	0.4	0.3	0.2	0.3	0.3	0.2	0.2	0.1	0.1
SE	0.4	0.4	0.4	0.3	0.2	0.7	0.8	1.0	1.0	1.0	0.8
SB	0.4	0.4	0.4	0.3	0.2	0.0	0.0	0.0	0.0	0.0	0.0
AK	0.4	0.4	0.4	0.2	0.1	0.3	0.2	0.3	0.4	0.4	0.4
BÖ	0.4	0.4	0.4	0.0	0.1	0.3	0.7	0.8	0.8	0.8	0.8
KV	0.4	0.4	0.5	0.5	0.2	0.2	0.1	0.4	0.4	0.4	0.4
E/M/E	0.4	0.4	0.4	0.3	0.2	0.3	0.5	0.5	0.5	0.3	0.3
NQ	0.4	0.4	0.3	0.0	0.3	0.6	0.7	0.7	0.7	0.5	0.4
Avg.	0.31	0.31	0.31	0.25	0.23	0.32	0.45	0.56	0.63	0.58	0.54
±stD	0.12	0.12	0.13	0.15	0.09	0.19	0.27	0.30	0.31	0.33	0.33