The Optimum Clustering Framework: Implementing the Cluster Hypothesis

Norbert Fuhr Marc Lechtenfeld Benno Stein Tim Gollub

University of Duisburg-Essen Bauhaus-Universität Weimar

Document clustering offers the potential of supporting users in interactive retrieval, especially when users have problems in specifying their information need precisely. In this paper, we present a theoretic foundation for optimum document clustering. Key idea is to base cluster analysis and evalutation on a set of queries, by defining documents as being similar if they are relevant to the same queries. Three components are essential within our optimum clustering framework, OCF: (1) a set of queries, (2) a probabilistic retrieval method, and (3) a document similarity metric.

After introducing an appropriate validity measure, we define optimum clustering with respect to the estimates of the relevance probability for the query-document pairs under consideration. Moreover, we show that well-known clustering methods are implicitly based on the three components, but that they use heuristic design decisions for some of them. We argue that with our framework more targeted research for developing better document clustering methods becomes possible. Experimental results demonstrate the potential of our considerations.

1 Introduction

The vast amount of research on IR methods deals with ad-hoc retrieval. However, from a user-oriented perspective, this is only one of several methods of information access. In [11] a faceted classification of information seeking strategies is presented, pointing out that retrieval is only possible if i) the access *mode* is specification and ii) the access *method* is searching (i.e. the IR system is able to process the user's specification). In many cases, users have problems in specifying their information need, thus the appropriate access mode will be recognition. Document clustering is a means for supporting users in these situations [30], since the user only has to *recognize* the cluster that fits best to her current information need. Thus, clustering can be helpful at all stages of a search (i.e. collection clustering as well as result clustering), offering the user the possibility to choose from a number of clusters instead of formulating a query.

In contrast to document classification, clustering also has the potential of offering multiple alternative clusterings at a time (similar to faceted search, see e.g. [24, 22, 64]). For example, when searching for papers on document clustering, some users might be interested in the clustering methods themselves, others might be looking at the document representation used, and some might want to know about the type of documents on which the methods have been tested.

In this paper, we describe a model that provides not only a theoretical basis for improving current clustering methods, but also defines a framework for extensions like e.g. multiple clusterings.

The Probability Ranking Principle (PRP) [50] forms the theoretic foundation for probabilistic retrieval. Before the formulation of the PRP, the development of new retrieval models was a purely heuristic task: researchers proposed a mathematical model (like e.g. vector space [52] or fuzzy logic [48]) and combined it with varying amounts of heuristics, in order to arrive at a retrieval model. The quality of this model could be verified only empirically by performing retrieval experiments with the few test collections available. If the experiments showed good performance, the model was deemed to be reasonable. With the PRP however, there suddenly was a theoretic justification for a certain type of models, stating that probabilistic retrieval models give optimum performance when results are ranked by decreasing probability of relevance. So good retrieval quality for probabilistic models is not a mere coincidence like with non-probabilistic ones – there is a theoretic proof that these models allow for optimum retrieval quality. However, as the PRP only provides a framework, there still remained the task of formulating actual models (based on different document and query representations plus using additional assumptions about the (in)dependence of certain types of events). In fact, it took almost 15 years until there were probabilistic models which outperformed the non-probabilistic ones in experimental evaluations [53, 18].

Looking at the field of document clustering, we are still in the pre-PRP era: although there is a vast amount of knowledge about the properties of various clustering methods, document clustering still is mainly based on heuristics with regard to the choice of document representation and similarity function. The quality of these choices can only be evaluated experimentally. There is no theoretic foundation which links the design of a document clustering method to the quality of its outcome.

In this paper we present a theoretic model which solves this problem: Similar to the formulation of the PRP, we will show how document representation and similarity function can be linked to the expected quality of the resulting clustering, thus providing a framework for optimum clustering. The development of actual clustering methods following this framework is *not* a subject of this paper—although we provide some experimental evidence indicating the validity of our approach.

The starting point for the development of our framework is the cluster hypothesis. According to it, cluster analysis could be used to support the identification of relevant documents given a request: Similar documents tend to be relevant to the same information need [60]. However, evaluations of the cluster hypothesis gave inconclusive results [61]. Instead, Hearst and Pedersen propose to use cluster analysis as a post-processing step on the set of retrieved documents. They argue that, especially for high dimensional data, various aspects of the data could serve as a basis for similarity assessment [23]. Which of these aspects are suitable to distinguish between relevant and irrelevant documents depends on the actual queries, and thus the (pre-) existence of a universally meaningful (= without considering knowledge about the information need) clustering cannot be expected. On the other hand, this argument also supports the idea of multiple alternative clusterings mentioned above.

In this work, we reverse the cluster hypothesis to motivate the introduction of a query set in combination with relevance assessments to improve the clustering of documents: *Documents relevant to the same queries should occur in the same cluster.* For achieving this, we redefine the concept of document similarity: Two documents are similar if they are relevant to the same queries. Thus, the original cluster hypothesis becomes a kind of tautology, since document similarity is no longer a property of its own, but depends on relevance. Moreover, by considering relevance, clustering is able to address the pragmatic level of information access, whereas the

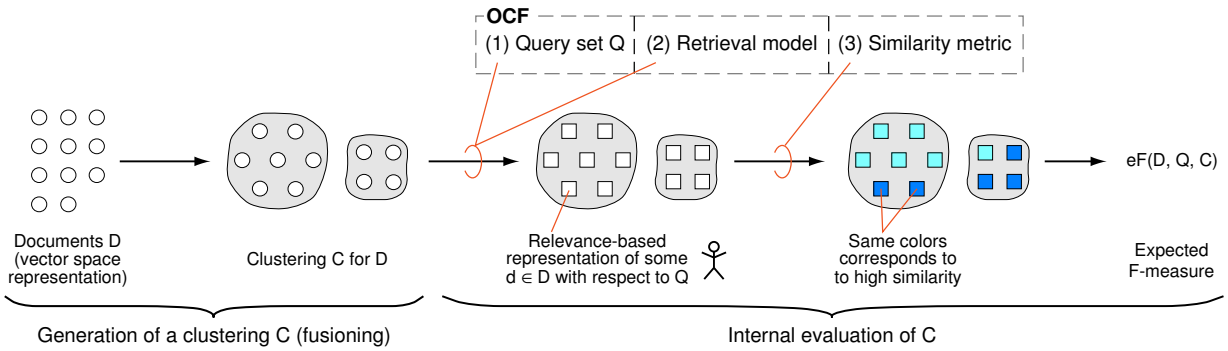

Figure 1: Illustration of the optimum clustering framework, OCF, here used as internal evaluation criterion to assess a given clustering $\mathcal{C}$. Salient property of OCF is an information-need-driven computation of document similarities: based on a query set Q relevance-based representations of the documents D are computed, and from the correlation of the resulting similarity graph and $\mathcal{C}$ the so-called "expected F-measure" is derived.

traditional view of document similarity is more at the semantic level.

Since we usually won't have explicit relevance information, we regard the probability of relevance instead. For supporting interactive information access, in the ideal case, the query set should contain an element that fits to the user's current information need, and the corresponding relevant documents should be clustered together. In fact, result clustering can be interpreted along these lines, where the query set consists of all possible refinements of the original query.

Looking at the standard approach to document clustering, documents are represented as bags of words (BOW). In the spirit of our idea, this approach uses the collection vocabulary as single term queries; thus, two documents are considered similar if they share a majority of words. This interpretation brings forth that the BOW method describes one arbitrary way to construct a query set and that different approaches stand to reason. Referring to the observations of Hearst and Pederson we believe that cluster analysis could be improved by applying a target-oriented set of queries (e.g. by following the idea of faceted search as mentioned above). Especially, any domain knowledge should be incorporated into the query set. Depending on the way domain knowledge is captured, different document models become related to our work (see Section 2).

The introduction of a query set tackles the clustering process at different stages. We therefore summarize our approach as the Optimum Clustering Framework (OCF—see Figure 1). As will be discussed in greater detail in Section 3 and 4, the term *optimum clustering* refers to a clustering that satisfies the (reversed) cluster hypothesis best. To be able to award this optimum property to a clustering, a validity metric based on relevance assessments is presented. Section 4 also describes how a query set Q could be operationalized to compute similarity scores between pairs of documents: Documents are represented through their probabilities of relevance estimated by a retrieval model for the query set Q. Similarity is then computed based on these representations, i.e., documents are not compared directly to each other but via their relevance to Q. In Section 5, a discussion on concrete query sets and their close relation to existing approaches is brought up. Section 6 deals with fusion principles that are tailored to the metrics of the OCF. To illustrate and substantiate the applicability of our framework, the design and the results of some experiments are presented in Section 7.

The major contribution of this paper is the development of a solid theoretical framework for future research on document clustering.

2 Related Work

Document clustering has a long history of research, see e.g. [60, 15]. The decisive point of our work is the introduction of relevance into the clustering process and indeed, early papers [26, 27] aimed at a closer connection between document similarity and relevance. But this line of research has not been continued. Robertson [50] even stated that the cluster hypothesis and the PRP are somewhat in contrast, since the former cannot be incorporated directly for computing the document-wise probabilities of relevance. However, as indicated above, our approach takes the opposite direction, using probabilities of relevance for generating clusterings.

More recent research has addressed the three major steps of document clustering: document representation [29], similarity computation [63, 41] and fusion [12, 33]. On the other hand, there has been an increasing number of applications of document clustering for various purposes. Besides the 'classic' approach of collection clustering—be it for supporting browsing (especially for topic detection and tracking [2]) or for cluster-based retrieval [61, 43], the focus has been mainly on result clustering, where the documents of the result set are grouped in order to structure the output [39]. In addition, these clusters can also be ranked [59, 37, 44, 36]. Other researchers have used result clustering for improving the ranked retrieval result, e.g. via cluster-based smoothing of documents [43, 14] or cluster-based resampling for pseudo-relevance feedback [35, 38].

In contrast to the vast amount of literature of document clustering methods, there are only a few user-oriented evaluations in this area. Some empirical user studies have shown that users prefer structured presentations of result sets over list-based ones [62] and help them in retrieving relevant documents. [30] found that result structuring supports vague information needs by making it possible for the user to use less precise queries and by forgiving mistakes in query formulation. The evaluations in [67, 23] showed that a topical structuring of the retrieval result set helps the user in getting an overview and understanding the content of the result set. The structuring also supports users in locating relevant documents more easily [67], especially when similar results from different locations are clustered within the ranked result list [30].

As mentioned in the opening section of this paper, the assessment of pairwise document similarities in the OCF also offers an information retrieval perspective to classify existing research. Dependent on how the query set is generated, the OCF evolves into pre-described approaches. To clarify this fact, we distinguish three paradigms of the query generation process: local, global, and extern.

1. Under the local paradigm the query set is generated by extracting words or phrases from each document in the collection independently. To assess the relevance of a document wrt. a query, traditional retrieval models like the vector space model, BM25 or language models are used. This way, the clustering process gets related to clustering under the bag-of-words model or keyword based clustering [42, 31].

2. Under the global paradigm queries are generated by considering global properties of the document set. Global properties may be topical or structural. Approaches to acquire topical queries and to assess their relevance wrt. a document include pLSI [25] or LDA [9]. Using structural queries, the OCF goes in line with work done in genre- or XML-clustering [56, 65].

3. Under the external paradigm query generation is based on any source of external knowledge. This may be in forms of manually created relevance judgments, user feedback, a foreign document collection like under the ESA (explicit semantic analysis) model [19], or even by operationalizing existing clusterings as done by semi-supervised clustering methods [13].

Note that, traditionally, most of the clustering approaches follow the local paradigm. The Optimum Clustering Framework can be applied under all paradigms, depending on the knowledge sources used to generate the query set Q.

In the next two sections, the idea of introducing a query set into the clustering process is formalized to an external and an internal validity measure. An illustration of how these measures are applied in a cluster analysis is given in Figure 1. Although our external measure relies on relevance judgments rather than on a reference classification, it is closely related to existing validity measures like the F-Measure [60, ch. 7] and the BCubed metric [5], which also consider a measure for precision and recall.

3 Cluster Validity under the OCF

Although there is a broad variety of cluster validity metrics, none of them is suited for our purpose: Whereas all popular metrics for external evaluation compare a clustering with a given document classification, we want to perform an evaluation with respect to a set of queries with relevance judgments—which is a generalization, since categories (and classification hierarchies as well) can be regarded as a special case of queries. In order to define optimum clustering and to be able to develop corresponding clustering methods, we need a metric satisfying the following two requirements:

1. The metric should be based on a given set of queries with complete relevance information.
2. It should be possible to compute expectations of this metric based on probabilistic retrieval models.

To derive our metric, some formal definitions are needed. Let $D = \{d_1, \ldots, d_N\}$ denote the set of documents to be clustered with respect to the query set $Q = \{q_1, \ldots, q_K\}$, where $\mathcal{R} \subset Q \times D$ denotes the set of relevant query-document pairs.

1 *For a document collection D, $\varphi_{\mathcal{C}} : D \to \mathbb{N}$ is a* clustering function *iff there exists some M such that $\varphi_{\mathcal{C}}$ is a surjective mapping into the interval $[1, M]$. Then we call the partitioning generated by $\varphi_{\mathcal{C}}$ a* clustering $\mathcal{C} = \{C_1, \ldots, C_M\}$ *with $C_i \subseteq D$ for $i = 1, \ldots, M$. Furthermore, we write $x \sim_{\mathcal{C}} y$ whenever $\varphi_{\mathcal{C}}(x) = \varphi_{\mathcal{C}}(y)$, and $x \not\sim_{\mathcal{C}} y$ otherwise*

2 *Two clusterings $\mathcal{C}$ and $\mathcal{C}'$ are* equivalent *iff $\forall (x, y) \in D \times D : x \sim_{\mathcal{C}} y \leftrightarrow x \sim_{\mathcal{C}'} y$. Let $D_R = \{x \in D | \exists q \in Q, \exists y \in D(x \neq y \wedge (q, x) \in \mathcal{R} \wedge (q, y) \in \mathcal{R})\}$ denote the* set of paired relevant documents. *Then two clusterings $\mathcal{C}$ and $\mathcal{C}'$ are* relevance-equivalent *iff $\forall (x, y) \in D_R \times D_R : x \sim_{\mathcal{C}} y \leftrightarrow x \sim_{\mathcal{C}'} y$.*

Furthermore, let $c_i = |C_i|$ denote the size of cluster C_i, and let $r_{ik} = r(C_i, q_k) = |\{d_m \in C_i | (q_k, d_m) \in \mathcal{R}\}|$ denote the number of relevant documents in C_i wrt. q_k.

Now we want to define a metric that reflects the cluster hypothesis. [28] already described a method for testing the clustering hypothesis by comparing document similarities of relevant-relevant vs. relevant-irrelevant document pairs. Besides reflecting the preferences of the cluster hypothesis, our metric also should allow for easy computation of expectations, so that the clustering process is able to target at optimum performance (like with the PRP mentioned in the beginning, enabling probabilistic retrieval methods to yield optimum retrieval quality).

The basic idea of our metric is, for each given query, to count the pairs of relevant documents occurring in the same cluster, and divide it by the total number of pairs in the cluster. Since we are focusing on pairs of relevant documents, unary clusters always get a value of 0—even if they

contain the only relevant document of a query. So we define our new measure *pairwise precision* P_p as the weighted average over all clusters:

$$P_p(D, Q, \mathcal{R}, \mathcal{C}) = \frac{1}{|D|} \sum_{\substack{C_i \in \mathcal{C} \\ c_i > 1}} c_i \sum_{q_k \in Q} \frac{r_{ik}(r_{ik} - 1)}{c_i(c_i - 1)} \tag{1}$$

As a simple example, assume that we have a disjoint classification with two classes a and b, and the documents are partitioned in three clusters: $(aab|bb|aa)$ (throughout this paper, we use this simplified notation for the case of a classification with disjoint classes as query set, where we denote only the classes of documents, and separate clusters by '|'). Then we would have $P_p = \frac{1}{7}(3(\frac{1}{3} + 0) + 2(0 + 1) + 2(1 + 0)) = \frac{5}{7}$. Note that we do not normalize by the number of queries—which is an arbitrary choice: This way, a perfect clustering for a disjoint classification (with $D_R = D$) will reach a P_p value of 1. For an arbitrary query set, however, we might also get values greater than 1, or the maximum value may be smaller than 1.

Like in retrieval, we need a second measure for considering all aspects of quality, which we call analogously *pairwise recall* R_p. For that, let $g_k = g(q_k) = |\{d \in D | (q_k, d) \in \mathcal{R}\}|$ denote the total number of relevant documents for q_k. Then we define

$$R_p(D, Q, \mathcal{R}, \mathcal{C}) = \frac{\sum_{q_k \in Q} \sum_{C_i \in \mathcal{C}} r_{ik}(r_{ik} - 1)}{\sum_{\substack{q_k \in Q \\ g_k > 1}} g_k(g_k - 1)} \tag{2}$$

For our three example clusters $(aab|bb|aa)$ from above, the clustering has produced four a pairs (out of twelve) and two b pairs (out of six), thus yielding $R_p = \frac{4+2}{12+6} = \frac{1}{3}$.

Equation (2) computes the so-called micro average by summing over all numerator and denominator values first, and then forming the quotient. This has the advantage that one can compute unbiased estimation values (by ignoring the denominator which is constant for a set of queries). Alternatively, one could also regard the macro average (arithmetic mean) over all queries with more than one relevant answer—but this solution would only allow for biased estimates.

In the following, we also regard the pairwise F-measure F_p which computes the harmonic mean of P_p and R_p:

$$F_p(D, Q, \mathcal{R}, \mathcal{C}) = \frac{2}{\frac{1}{P_p(D,Q,\mathcal{R},\mathcal{C})} + \frac{1}{R_p(D,Q,\mathcal{R},\mathcal{C})}} \tag{3}$$

The metric proposed here is similar to other pairwise measures, especially the Folkes and Mallows FM metric (see e.g. [3]) in that both approaches compute recall and precision wrt. item pairs. But, the FM measure computes the geometric mean of these, whereas we keep the two metrics separate, and prefer the F measure for combining them for the reasons described in [60, ch. 7]. The BCubed metric [5] also computes recall and precision, but averages on a per item basis. Furthermore, both FM and BCubed have been defined for disjoint categories only. The Rand index [49] is a single measure regarding object pairs. However, all these metrics suffer from the problem that a pair of irrelevant documents counts as much as a pair of relevant ones—which contradicts our standpoint. Second, computation of unbiased expectations is difficult to impossible in most cases. Finally, having two metrics instead of a single one gives more flexibility wrt. user preferences, like in ad-hoc retrieval.

We point out that the F_p measure possesses properties desirable for cluster metrics in general. In the appendix, we show that the F_p measure satisfies the four axioms of Ackerman & Ben-David, which are intended to capture the essence of reasonable clustering quality measures [1]. The development of these axioms is inspired by Kleinberg's three axioms for clustering functions [34],

which formalize a set of properties that appear plausible at first sight, but which are inconsistent: Kleinberg's 'impossibility theorem for clustering' [34] states that no clustering function (= fusion principle) can exist that fulfills all axions at the same time. However, in [66] a different variant of Kleinberg's axioms is suggested, proving that single-link is the only clustering method satisfying their set of axioms. Altogether, we consider the work of [1] relevant for us: it focuses on clustering quality measures (in the authors' sense synonymous with internal validity measure), this way asking for what an optimum clustering constitutes, and not how to construct it. Finally, a more descriptive set of constraints for clustering metrics is proposed in [3], which we also discuss in the appendix.

4 From Perfect to Optimum Clustering

Given the two measures as described above, we can now define perfect and optimum clustering. We chose the terminology analogously to classic retrieval: In *perfect retrieval*, all relevant documents are ranked ahead of the first non-relevant one. However, since a real IR system has only limited knowledge about information needs and the meaning of documents, it operates on representations of these objects, and can only achieve *optimum retrieval* with regard to the representations.

In the following, we first define the perfect variant, and then introduce optimum clustering. For this discussion, we focus on classes of relevance-equivalent clusterings, since the distribution of the documents from $D-D_R$ will not affect the values of P_p and R_p, as long as these documents are kept separate from those in D_R. Below, we shortly talk about classes of clusterings when we are referring to the equivalence classes wrt. relevance equivalence.

3 *For a document collection D, a set of queries Q and a corresponding relevance relation $\mathcal{R} \subseteq Q \times D$, $\mathcal{C}$ is a perfect clustering iff there exists no clustering $\mathcal{C}'$ of D with $P_p(D,Q,\mathcal{R},\mathcal{C}) < P_p(D,Q,\mathcal{R},\mathcal{C}') \wedge R_p(D,Q,\mathcal{R},\mathcal{C}) \leq R_p(D,Q,\mathcal{R},\mathcal{C}')$ or $P_p(D,Q,\mathcal{R},\mathcal{C}) \leq P_p(D,Q,\mathcal{R},\mathcal{C}') \wedge R_p(D,Q,\mathcal{R},\mathcal{C}) < R_p(D,Q,\mathcal{R},\mathcal{C}')$.*

Our definition of perfect clustering is a strong Pareto optimum (see e.g. [16]). As for Pareto optima in general, there usually will be more than one perfect clustering for a given triple $(D,Q,\mathcal{R})$. The case of a disjoint classification is an exception, as well as any triple where no document is relevant to more than one query.

As a simple example, assume that we have five documents and two queries where d_1, d_2, d_3 are relevant for q_1, and d_3, d_4, d_5 are relevant for q_2. Here we get the highest precision ($P_p = 1$) if we cluster either the first three or the last three documents in one cluster and the other two in another cluster, along with a recall of $R_p = \frac{6+2}{6+6} = \frac{2}{3}$. Thus, even for a single Pareto optimum (R_p, P_p), there may be more than one corresponding clustering (which is good news, increasing the chances of a clustering algorithm of finding at least one of these). The other Pareto optimum in this example results from putting all documents in one cluster, yielding the maximum recall ($R_p = 1$), but a precision of $P_p = 0.6$, only.

Unfortunately, the set of perfect clusterings is not even guaranteed to form a cluster hierarchy, as can be easily seen from the following example: Let there be 4 documents $d_1, \ldots, d_4$ and 4 queries a, b, c, d, where the documents are relevant to the following queries: $d_1 : \{a, c\}$, $d_2 : \{a, b\}$, $d_3 : \{b, c, d\}$, $d_4 : \{d\}$. Now we can either cluster $\mathcal{C} = \{\{d_1, d_2, d_3\}, \{d_4\}\}$ or $\mathcal{C}' = \{\{d_1, d_2\}, \{d_3, d_4\}\}$. Then we have $P_p(\mathcal{C}) = \frac{3}{4}$ and $R_p(\mathcal{C}) = \frac{3}{4}$ vs. $P_p(\mathcal{C}') = 1$ and $R_p(\mathcal{C}') = \frac{1}{2}$ (forming a single cluster would yield $P_p = \frac{2}{3}$ only). Thus, in general, the set of perfect clusterings cannot be generated

by a hierarchic clustering algorithm. Even with complete knowledge of $(D, Q, \mathcal{R})$, any hierarchic method is bound to miss some of the Pareto optima.

Now we turn to optimum retrieval, which we base on expected values of pairwise recall and precision. Thus, we switch from measures for external evaluation to internal ones. For that, we assume that we have a retrieval method which is able to estimate the probability of relevance $P(rel|q,d)$ of a given query-document pair (q,d). Then, for each document pair, we can estimate the probability that both documents are relevant, by assuming that the probabilities of relevance of the two documents are independent. This seems to be counter-intuitive in the view of classical clustering approaches, where the relevance probabilities of two similar documents cannot be regarded as being independent. However, in our approach, documents are not similar per se, similarity is defined only on top of their retrieval scores for the given query set, as we will see below. Thus, there is no contradiction to the independence assumption.

From the probabilities of two documents being both relevant, we can sum up over all document pairs in order to get an unbiased estimate of the number of relevant document pairs in a cluster (note that estimating this value from the expected number of relevant documents $E(r_{ik})$ would result in a biased estimate). In order to simplify the following definitions, we introduce the *expected cluster precision* (ecp). First, we define *restricted ecp* $\tilde{\sigma}(C)$ for clusters C with at least two elements:

$$\begin{aligned}\tilde{\sigma}(C) &= \frac{1}{c(c-1)} \sum_{q_k \in Q} \sum_{\substack{(d_l,d_m)\in C\times C \\ d_l \neq d_m}} P(rel|q_k,d_l)P(rel|q_k,d_m) \\ &= \frac{1}{c(c-1)} \sum_{\substack{(d_l,d_m)\in C\times C \\ d_l \neq d_m}} \sum_{q_k \in Q} P(rel|q_k,d_l)P(rel|q_k,d_m) \qquad (4)\end{aligned}$$

For clusters with only one element, we define an ecp value of 0 (as for P_p). We can now define ecp for all cluster sizes as $\sigma(C) = \tilde{\sigma}(C)$, if $|C| > 1$, and $\sigma(C) = 0$, otherwise.

As eqn (4) shows, we regard each document pair wrt. all queries. Thus, for each document, we are only interested in its probability estimates for the given query set, and so we can transform a document into a vector of relevance probabilities $\vec{\tau} : D \to [0,1]^{|Q|}$ with $\vec{\tau}^T(d_m) = (P(rel|q_1,d_m), P(rel|q_2,d_m), \ldots, P(rel|q_{|Q|},d_m))$. With this notation, we can express the restricted ecp as follows:

$$\tilde{\sigma}(C) = \frac{1}{c(c-1)} \sum_{\substack{(d_l,d_m)\in C\times C \\ d_l \neq d_m}} \vec{\tau}^T(d_l) \cdot \vec{\tau}(d_m) \qquad (5)$$

Based on these definitions, we can now estimate the quality of a clustering. Expected precision can be computed as the weighted average (considering cluster size) of the clusters' ecp values:

$$\begin{aligned}\pi(D,Q,\mathcal{C}) &= \frac{1}{|D|} \sum_{\substack{C_i \in \mathcal{C} \\ |C_i|>1}} \frac{1}{c_i - 1} \sum_{\substack{(d_l,d_m)\in C_i\times C_i \\ d_l \neq d_m}} \vec{\tau}^T(d_l) \cdot \vec{\tau}(d_m) \qquad (6) \\ &= \frac{1}{|D|} \sum_{C_i \in \mathcal{C}} c_i \sigma(C_i) \qquad (7)\end{aligned}$$

For expected recall, a direct estimation would lead to the problem that we would also have to estimate the denominator, which would result in a biased estimate of the recall. However,

since the denominator is constant for a given query set, we can ignore this factor, as we are only interested in comparing the quality of different clusterings for the same (D, Q) pair. So we omit the denominator and compute an estimate for the numerator only:

$$\rho(D,Q,\mathcal{C}) = \sum_{C_i \in \mathcal{C}} \sum_{\substack{(d_l,d_m) \in C_i \times C_i \\ d_l \neq d_m}} \vec{\tau}^T(d_l) \cdot \vec{\tau}(d_m) \tag{8}$$

$$= \sum_{C_i \in \mathcal{C}} c_i(c_i - 1)\sigma(C_i) \tag{9}$$

Based on these definitions, we also define the *expected F-measure* as the harmonic mean of π and ρ:

$$eF(D,Q,\mathcal{C}) = \frac{2}{\frac{1}{\pi(D,Q,\mathcal{C})} + \frac{1}{\rho(D,Q,\mathcal{C})}} \tag{10}$$

With these metrics, we can now introduce optimum clustering:

4 *For a document collection D, a set of queries Q and a retrieval function yielding estimates of the probability of relevance $P(rel|q,d)$ for every query-document pair (q,d), $\mathcal{C}$ is an optimum clustering iff there exists no clustering $\mathcal{C}'$ of D with $\pi(D,Q,\mathcal{C}) < \pi(D,Q,\mathcal{C}') \wedge \rho(D,Q,\mathcal{C}) \leq \rho(D,Q,\mathcal{C}')$ or $\pi(D,Q,\mathcal{C}) \leq \pi(D,Q,\mathcal{C}') \wedge \rho(D,Q,\mathcal{C}) < \rho(D,Q,\mathcal{C}')$.*

The major difference to the definition of perfect retrieval lies in the replacement of the actual relevance judgments by the estimations of the probability of relevance.

As with perfect clustering, we are targeting at a set of strong Pareto optima. Due to the uncertain knowledge, we usually have more optimum than perfect solutions. This is no surprise, taking into account that we are optimizing wrt. a set of queries: For a single query, there is only one perfect retrieval result (retrieving all and only relevant documents), but optimum retrieval can only be performed in form of a ranking. When we are asking for hard clustering, then it is impossible to find a single optimum solution.

With the above definition, one can compare different cluster distributions of a document collection, in order to find an optimum clustering. Thus, a brute force clustering algorithm would work as follows:

1. For a given document collection, a set of queries and a probabilistic retrieval method, generate all possible clusterings, and compute expected recall and precision for each clustering.

2. Determine those clusterings fulfilling the condition for optimum clustering.

Of course, due to the exponential number of possible clusterings of a document set, this approach is not feasible—but see the discussion on existing fusion principles in Section 6 and the experiments in Section 7. In the next section, we describe how the components that are needed for an evaluation of clusterings with the OCF can be designed.

5 Choosing the OCF Components

In our view, all document clustering methods rely on three components:

1. a set of queries,

2. a retrieval function, and

3. a document similarity metric.

The standard clustering approach (representing documents as bags of words) uses the collection vocabulary as single term queries. Other clustering approaches that are based on more advanced document representations use the document features in the collection (or a subset thereof) as queries. In addition, a clustering method defines a weighting function for each feature in a document, which reflects the importance of the feature wrt. the document. In our view, this weight represents the retrieval score of the document for a query with the specific feature. Finally, the document similarity metric employed by most approaches is the cosine measure or the scalar product.

The OCF approach covers most (if not all) of the existing document clustering methods. These methods are often defined as heuristics, without a theoretic foundation that links the choice of the three components to the resulting clustering quality. Given the OCF, we have a foundation, and thus, a more targeted development towards better clustering methods is possible. At first glance, the OCF seems to be more complicated than the existing methods, e.g. by requiring a probabilistic weighting function. On the other hand, the existing methods can be interpreted as heuristic approximations to our framework, and by referring to the OCF, we immediately see possibilities for improvements, for example: Are there better query sets than single terms? Can we replace the $tf \cdot idf$ weighting by a probabilistic one? In general, one should choose the components in such a way that they fit to the underlying theoretic model. As we will show below, there is no degree of freedom in the choice of the similarity metric, and little in the definition of the retrieval function. For the query set, the OCF poses no restrictions. However, our approach of interpreting the document features as relevance assessments gives us a strong hint on how to form a query set: In the ideal case, queries should capture the users' information needs—then applying the OCF will cluster relevant documents together. In the following, we discuss each of the three components in greater detail and demonstrate a number of design options available for each of them.

Query Set The main challenge of query set generation is to find queries related to the users' current information needs. Ideally, the system would have context information about the user and her information need, thus being able to generate a context-specific query set, which would lead to context-specific clustering. As mentioned before, result clustering (e.g. [39]) can be regarded as an approach along these lines, where the query set consists of possible refinements of the current query. For collection clustering, three paradigms for query set generation have been used, as outlined in Section 2. Whereas the local and global methods find a query set by analyzing the given document collection, external methods focus on the incorporation of domain knowledge or common sense knowledge.

The simplest local method for query set generation is to use each unigram that occurs in the document collection as one query. Consequently, a document will be represented by its relevance wrt. each unigram. This method resembles clustering under the bag-of-words model. As an alternative local approach, keyphrases can be extracted from the documents to form a more focused query set. See Section 7 for a demonstration of this approach. In our ongoing research, we also investigate whether stylometric features like readability indexes or POS-features can contribute to a more sophisticated query set. Besides the standard term-based models, more advanced topic models operating on transformations of the term space are also feasible. These methods follow the global paradigm. There is already a range of clustering methods based on latent variables (see e.g. [8, 21, 25]); in terms of our framework, the single dimensions or concepts resulting from these transformations form query candidates.

The most promising methods for query set generation are those driven by external resources. An existing approach following the external paradigm is the ESA model. In the perspective of our framework, ESA uses a set of Wikipedia articles as query set. It is noticeable how natural the model evolves from the idea behind the OCF. Besides Wikipedia, there are plenty of other resources that could serve as a basis for external query set generation: tags of a social tagging system, titles from news portals, or queries from a query log, to name a few. A crucial point is that external resource should reflect the user's raw notions of the document collection. A definite guide is still missing here, but we believe that the use of external resources is a fruitful field for further research. A maybe more directed approach in this respect is to engage human reviewers to study parts of the document collection and to create an initial query set manually. The scalability of this idea is limited, but the emergence of crowdsourcing platforms like Amazon's mechanical turk [1] show its feasibility.

There also is the potential for using multiple query sets in order to generate multiple clusterings from which the user can choose—either for the whole collection or context-dependent as a form of result clustering. In our view, faceted search [24, 22, 64] can be regarded as a set of clusterings along different dimensions, which are usually defined via formal attributes. We suggest to generalize this idea towards arbitrary facets. One possible source for defining these facets is the document structure: emails can be clustered by subject, by sender or by date, patent documents by claims, previous work or by the description of the innovation; descriptions in online bookstores can be grouped by editorial reviews, user reviews, or even book covers. When documents are unstructured, external resources can provide the necessary information for generating different query sets. By generating multiple clusterings, the search interactions are richer, even if no formal criteria for performing standard faceted search are available, and so interactive retrieval becomes more effective [17].

Probabilistic Retrieval Function The design decisions made to construct the query set sometimes prescribe a specific retrieval function, as with latent variable models or stylometric features, for example. Often, however, there is a broad variety of retrieval functions one can choose from, e.g. $tf \cdot idf$, BM25, or language models. An arising problem is the estimation of the actual probability of relevance, since most retrieval methods compute a score that is only rank-equivalent to probabilistic retrieval. In most cases, certain query-specific constants are ignored within the score computation. However, our framework heavily relies on the comparison of probabilities of relevance for different queries (e.g. are the documents $\{d_1, d_2\}$ more likely to be relevant to q_1 than are $\{d_2, d_3\}$ to q_2?). Besides trying to apply the underlying models in such a way that they directly estimate the probability of relevance, there are also methods for transforming the retrieval score into such a probability; examples include [47] or the recent studies on score distributions [4, 32].

Document Similarity Metric Regarding the addends of eqn (5), the obvious measure for document similarity is the scalar product of the $\vec{\tau}(d)$ vectors, which yields the expected number of queries for which both documents are relevant. In case the $\vec{\tau}$ vectors contain $tf \cdot idf$ weights, we have a standard similarity metric, for which our framework gives a new interpretation. We like to point out that the relationship between the similarity metric and the cluster hypothesis has been investigated by previous researchers: whereas [61] used standard document similarity measures for testing the cluster hypothesis, [58] and [54] investigated the idea of using query-specific similarity metrics by combining the standard document similarity with query-specific weights.

[1] http://aws.amazon.com/mturk/

While the latter two papers consider the actual query to a certain extent, our similarity metric is based on a set of queries.

6 Existing Fusion Principles in the Light of the OCF

Across the broad range of fusion principles that have been developed in the past are those most amenable to our framework that analyze the cluster quality after each fusion step (agglomerative methods) or division step (divisive methods). Running such a method under OCF simply means to employ one or both of the two metrics expected cluster precision (eqn (5)) and expected recall (eqn (8)) as quality measure. Under an agglomerative clustering method the former corresponds to the metric employed in group average clustering [55], where also all pairs of the resulting cluster are considered. Each step of this method results in a cluster with higher (or equal) recall ρ than the two clusters being merged. By contrast, expected precision π will decrease (ignoring the unary clusters with $\pi = 0$ here), as can be easily seen from its definition as the average similarity of the $\vec{\tau}$ vectors. Divisive methods, on the other hand, start with a single cluster which has maximum expected recall, but very low precision. Then they divide clusters in order to increase precision, but with minimum loss in recall. Among these methods, min-cut [46] comes closest to the OCF: Considering the scalar product of the $\vec{\tau}$ vectors as similarity and thus as edge weight in the similarity graph, the min-cut criterion corresponds to finding those edges that minimize the reduction in expected recall, as can be seen from eqn (8). For breaking ties, the OCF suggests to consider the expected pairwise precision of the result of the divisive step.

Although we have shown that no hierarchical clustering method is able to find all Pareto optima, it is an interesting question whether or not at least one optimum is reached. In fact, we can show that min-cut finds such an optimum: First let us assume, that the similarity graph is cohesive (if not, min-cut will first divide it into its cohesive subgraphs, thus increasing precision without reducing recall). Then we have the first optimum with maximum expected recall. In the next step, min-cut will search for a cut with minimum reduction of recall. If there is more than one decomposition possibility with the same reduction in recall, then we should choose the one with the highest expected precision. This way, we will arrive at the next Pareto optimum with the second highest recall and maximum precision for this recall point. For the next min-cut step, however, we cannot show the same property, since a better result might be reached via a suboptimal choice in the first cut step (see e.g. the example with 4 documents and 4 queries from Section 4). This result is theoretically interesting, since there are $2^{N-1}-1$ nontrivial possibilities for bipartitioning a collection of N documents, and for some clustering criteria, finding the optimum solution is NP-hard [20]. In contrast, for our criterion, min-cut finds the optimum solution in $O(N^3)$ steps. From a practical point of view, however, this result is of little value, since we usually want a much larger number of clusters with a significantly higher precision.

For the agglomerative methods, we cannot show that they will find an optimum. This is due to the definition of pairwise precision, where any singleton cluster has a precision of 0. Thus, in general, we will need several fusion steps in order to reach the first optimum – and a greedy strategy will hardly find these steps.

Overall, the choice of the new quality measures pairwise recall and precision seems to be well justified, since it gives a posteriori nice theoretical foundations of some cluster similarity metrics and fusion principles, and also highlights their properties wrt. optimum clustering.

Table 1: Overview of the experimental design.

Variable	Variation	Effectiveness analysis	
Collection	Size Number of classes Class distribution		**Tables 2+3**
Fusion principle	Iterative: k-means Hierarchical: group average Random assignment	Dependency on collection (F-measure)	**Table 4**
Internal validity measure	Distance: Dunn-Index Shape: silhouette Graph: expected density OCF: terms, keywords	Dependency on collection and fusion principle (correlation coefficient)	**Table 5**

7 Experimental Analysis

The OCF provides a new way to model similarity in text clustering applications. If clustering is understood as a method to group documents according to query relevance, the OCF explains also at which places domain knowledge should be integrated. What are the practical impacts, in terms of new clustering technology, that can be expected from the OCF? In this section, we answer this question and report on appropriate experiments.

Document clustering technology combines the three OCF elements (document representation, retrieval function, similarity measure) with a fusion principle. Popular methods use the bag of words approach in combination with the vector space model or a probabilistic model and cosine similarity as OCF elements. With respect to fusion principles, the picture is less clear; the iterative principle in the form of k-means is pretty popular, but the hierarchical or the density-based principle work better in most applications. As mentioned before, the OCF can be combined with any fusion principle.

In order to demonstrate the validity of the OCF, we compare the OCF quality metric with existing metrics, both qualitatively and consistently over a relevant application range. For that we analyze the effectiveness of the expected F-measure when being used as *internal validity measure*, i.e., when being used for the discrimination between good and bad clusterings of a data set. We would like to point out the utmost importance of internal validity measures, as they finally decide about clustering quality: in practical applications several alternative clusterings are generated amongst which the best one has to be chosen—without having access to the ground truth (a crucial fact which is often neglected in experimental analyses).

We devise the experimental setup shown in Table 1, addressing the following issues of the expected F-measure metric:

1. Robustness with respect to the number of classes and class imbalance.
2. Independence of fusion principles.
3. Independence of high fusion variance.
4. Capturing of distance, shape and density characteristics.
5. Positive correlation with respect to query term expressiveness.

Table 2: Statistics of the five test collections (RCV subsets) used within the experiments.

Collection	Size	Number of classes	Class distribution	Classes
rcv-ss1	600	6	uniform	C11, E131, GCAT, GSCI, GSPO, M143
rcv-ss2	6 000	6	uniform	·/·
rcv-ss3	9 000	6	uniform	
rcv-ss4	6 000	6	2 000, 2×1 500, 500, 2×250	(corresponding to class sequence above)
rcv-ss5	6 000	12	uniform	C11, C23, C311, E71, E131, GCAT, GCRIM, GSCI, GSPO, GVIO, M11, M143

The main results of our analysis are comprised in Table 5. In addition, Table 4 reports on the clusterability of the test collections against the algorithms k-means, group average, and random assignment; this information is helpful in answering the issues 2 and 3. We detail the experimental setup in the following.

Five different subsets of the Reuters news corpus RCV1 are used as test collections [40]. The documents are drawn iid. from up to 12 classes, covering the four top-level classes of the RCV1 while considering only documents that are labeled with exactly one class. Tables 2 and 3 describe the collections and the used classes.

Table 3: Description of the classes in the five test collections (RCV subsets).

Class	Description	Used	Parent
CCAT	Corporate / Industrial		Root
C11	Strategy / Plans	*	
C23	Research / Development	*	CCAT
C31	Markets / Marketing		
C311	Domestic markets	*	C31
GCAT	Government / Social	*	Root
GCRIM	Crime, Law enforcement	*	
GSCI	Science and technology	*	GCAT
GSPO	Sports	*	
GVIO	War, Civil war	*	
ECAT	Economics		Root
E71	Leading indicators	*	ECAT
E13	Inflation / Prices		
E131	Consumer prices	*	E13
MCAT	Markets		Root
M11	Equity markets	*	MCAT
M14	Commodity markets		
M143	Energy markets	*	M14

Within all experiments, documents are represented under the bag of words model, applying the BM25 weighting function with the suggested standard parameters[2] to compute the weight of

[2] "Experiments have shown reasonable values are to set k_1 (and k_3) to a value between 1.2 and 2 and $b = 0.75$." [45]

each term t for the vector representation of document d:

$$w(t,d) = \mathrm{idf}(t) \cdot \frac{(k_1+1)\cdot \mathrm{tf}(t,d)}{\mathrm{tf}(t,d) + k_1 \cdot (1 - b + b \cdot \frac{L_d}{L_{avg}})} \ ,$$

where $\mathrm{idf}(t)$ denotes the inverse document frequency of term t in the collection, $\mathrm{tf}(t,d)$ the term frequency of term t in document d, L_d the length (number of words) of d, and L_{avg} the average length of a document in the collection.

Table 4: External evaluation of fusion principles with respect to their effectiveness of class identification. The table shows both mean μ and standard deviation σ of the achieved F-measure values over 40 clustering runs per each combination of collection and fusion principle.

Fusion	**rcv-ss1**		**rcv-ss2**		**rcv-ss3**		**rcv-ss4**		**rcv-ss5**	
principle	F: μ	σ	F: μ	σ	F: μ	σ	F: μ	σ	F: μ	σ
k-means	0.47	0.11	0.54	0.08	0.53	0.08	0.60	0.10	0.36	0.06
Group average	0.60	0.11	0.66	0.10	0.69	0.10	0.52	0.10	0.54	0.15
Random assignment	0.58	0.24	0.60	0.24	0.59	0.24	0.63	0.22	0.55	0.27

The cluster algorithms k-means (iterative fusion principle) and group average (hierarchical fusion principle) apply the cosine similarity to compute the pairwise similarities between the BM25 document vectors. For both algorithms the true category number is withhold. For k-means the parameter k is varied between 2 and 41, while for group average 40 clusterings with 2, 4, ..., 80 clusters are taken as defined by the dendrogram. These generation strategies account for the different sensitivity of the two fusion principles. While k-means often leads to very different clusterings when varying k, two consecutive clusterings of group average differ only by the fusion of two clusters with minimum distance. Under the random assignment algorithm, 40 clusterings are generated such that in clustering i ($i = 1, \ldots, 40$), a document is assigned to its correct class with probability $1/i$ and with probability $1 - 1/i$ randomly to one of the other classes (i.e. $i = 1$ yields a perfect clustering, while $i = 40$ results in the worst one). Altogether, $5 \cdot 3 \cdot 40 = 600$ clusterings are computed. Table 4 reports the mean and the standard deviation of $F(\mathcal{C}, \mathcal{C}^*)$, the achieved F-measure values, which are computed under the set-matching paradigm:[3]

$$F(\mathcal{C}, \mathcal{C}^*) = \sum_{c_j^* \in \mathcal{C}^*} \frac{|c_j^*|}{n} \max_{c_i \in \mathcal{C}} \left\{ F(c_i, c_j^*) \right\}$$

Here n is the number of documents in the collection, while $\mathcal{C}$ and $\mathcal{C}^*$ denote the clustering and the true classification respectively. $F(c_i, c_j^*)$ computes the standard F-measure for cluster c_i and true class c_j^* with equally weighted precision and recall.

From a clustering perspective, i.e., with respect to the effectiveness and its variance, the results in Table 4 are reasonable and in accordance with the literature [10][57]. Regarding our experiment design, however, we are not interested in the absolute performance of an algorithm but in whether we are able to spot a clustering with maximum F-measure within a set of clusterings. I.e., we are looking for an internal evaluation measure that is able to predict the ranking of the (true) externally computed evaluation results. In this regard Table 4 ensures that a broad range of clustering situations is considered.

[3] Under the counting-pairs paradigm the *document pairs* form the basis for the computation of precision and recall, entailing a different computation rule for $F(\mathcal{C}, \mathcal{C}^*)$ [3].

Table 5: Evaluation of internal validity measures with respect to their effectiveness of clustering selection. The table shows the achieved correlation coefficient values; evaluation basis are the 600 clustering experiments reported in Table 4.

Collection	**Fusion principle**	**Dunn-Index** ρ	**Silhouette** ρ	**Expected density** ρ	**OCF terms** ρ	**OCF keywords** ρ
rcv-ss1	k-means	0.38	0.53	0.65	0.77	**0.80**
rcv-ss2		0.67	-0.26	0.59	0.67	**0.90**
rcv-ss3		0.56	-0.31	0.72	0.78	**0.93**
rcv-ss4		0.61	-0.02	0.63	0.68	**0.81**
rcv-ss5		0.83	-0.48	0.85	**0.88**	0.83
rcv-ss1	Group average	0.21	-0.17	0.34	0.41	**0.61**
rcv-ss2		0.73	-0.61	0.82	0.85	**0.89**
rcv-ss3		0.79	-0.68	0.84	0.88	**0.90**
rcv-ss4		0.65	-0.44	0.66	0.70	**0.71**
rcv-ss5		0.70	-0.97	0.88	0.90	**0.98**
rcv-ss1	Random assignment	0.55	0.98	0.97	0.97	0.97
rcv-ss2		0.50	0.96	0.97	0.97	0.97
rcv-ss3		0.65	0.96	0.97	0.97	0.97
rcv-ss4		-0.22	0.97	0.97	0.97	0.97
rcv-ss5		0.64	0.98	0.97	0.97	0.97

Table 5 compares the effectiveness of five internal evaluation measures: the well-accepted Dunn-Index [7], which balances between inter- and intra-cluster distances, the shape-based silhouette coefficient [51], the graph-based measure expected density [57], and two new measures that apply the expected F-measure derived in our framework, called OCF terms and OCF keywords. The latter two measures differ in the specificity of the employed query set (explained below). The table quantifies the correlations between the rankings obtained from the internal measures and the true rankings, according to the Pearson correlation coefficient. The Dunn-index shows a consistent characteristic and yields acceptable results for the larger collections; its unsatisfying performance under random assignment is rooted in the fact that Dunn focuses on extremal similarities instead of averaging over all values: a local "decontamination" of an otherwise homogeneous cluster, which is likely under random assignment, is overrated. By contrast, the silhouette coefficient yields reasonable values only for clusterings generated by the random assignment algorithm. Silhouette prefers clusterings with few clusters, and a deeper analysis of the generated clusterings revealed that the best clusterings of k-means and group average typically contain many clusters. The expected density outperforms both the Dunn-index and the silhouette coefficient—a fact which has been reported before, for a variety of settings. The OCF-based measures perform best, where OCF keywords yields better values than OCF terms within all except one setting.

OCF-based measures quantify document similarity indirectly, via a query set Q. In particular, OCF terms considers all words of a collection's vocabulary as query set Q_T, which hence can be considered as a canonical query set. OCF keywords introduces more "query semantics" by extracting keyphrases from the documents of a collection, each forming an element of the query set Q_K. We use the method from Barker & Cornacchia for this purpose [6]: first, by a parts of speech analysis, the document is skimmed for base noun phrases. In a second step, scores are assigned to the extracted phrases based on the noun frequency and the phrase length. Third, single letter phrases and wholly contained subphrases are removed.

The relevance probability of a query $q \in Q$ wrt. a document d defines one component (di-

mension) in the vector $\vec{\tau}_Q(d)$ of relevance probabilities. This relevance probability is computed according to the BM25 retrieval model. The scalar product of two vectors $\vec{\tau}_Q(d_1)$ and $\vec{\tau}_Q(d_2)$ of relevance probabilities in turn defines the similarity between the two documents d_1 and d_2, and is accounted by the expected F-measure for a given clustering $\mathcal{C}$ (see eqn 5 to 10). With regard to the two query sets used, we have $|Q_K| < |Q_T|$. Our experiments show that the OCF-based measure benefits from focusing on the smaller keyphrase query set Q_K: the similarity graph computed for Q_K contains less noise and thus allows for a more stable validation. It may be interesting to learn that if the set Q_K is also used as vocabulary for representing the documents D, the quality of the found clusterings does not improve in terms of the external F-measure (we have repeated all experiments with the respective setting). I.e., while the OCF-based validity measure benefits from focusing on a small set of discriminating keyphrases, the clustering algorithms k-means and group average require a more fine-grained representation of the documents to inform their fusion process. Altogether, the most important insight relates to the possibility to introduce semantics into a clustering process, which may relate to user preferences or to task specifics—both can be expressed in query form. Also note that query sets as they are used within the OCF enable one to transfer clustering preferences between different document collections.

8 Conclusion and Outlook

We have devised a new framework for document clustering (OCF). As pointed out, any clustering method is based on a set of queries, a retrieval function and a document similarity metric. The document clustering methods developed in the past disregard or implicitly constrain these three essential components. In contrast, the OCF establishes a theoretic basis that tells us how we can achieve optimum clustering quality for a given query set and a probabilistic retrieval function, thus enabling more targeted research for developing better document clustering methods.

The results of our initial experiments demonstrate the potential of our approach, but are intended for illustrative purposes in the first place. Altogether, our major contribution is its well-founded theoretical framework for document clustering, in order to replace the currently prevailing heuristic methods by more solid approaches, just like the Probability Ranking Principle for probabilistic retrieval. We do not claim that we have already better clustering methods—this will be the issue of further research, for which we have laid the ground here.

The discussion in this paper has focused on hard clustering, and cluster hierarchies have only been addressed briefly. Extending our framework to soft clustering and/or cluster hierarchies is also a subject of further research.

As mentioned in the very beginning, we see clustering methods as an important tool for interactive information access. However, more user-oriented research is required for investigating the full potential of this concept.

Acknowledgment

This work was supported in part by the German Science Foundation (DFG) under grants FU205/22-1 and STE1019/2-1.

References

[1] M. Ackerman and S. Ben-David. Measures of clustering quality: A working set of axioms. In *Proceedings NIPS 200*, pages 121–128. MIT Press, 2008.

[2] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang. Topic detection and tracking pilot study final report. In *In Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, pages 194–218, 1998.

[3] E. Amigo, J. Gonzalo, J. Artiles, and F. Verdejo. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, 12(4):461–486, 2009.

[4] A. T. Arampatzis, S. Robertson, and J. Kamps. Score distributions in information retrieval. In *ICTIR '09*, pages 139–151. Springer, 2009.

[5] A. Bagga and B. Baldwin. Entity-based cross-document coreferencing using the vector space model. In *COLING-ACL*, pages 79–85, 1998.

[6] K. Barker and N. Cornacchia. Using noun phrase heads to extract document keyphrases. In *Proc. AI '00*, pages 40–52, London, UK, 2000. Springer-Verlag.

[7] J. Bezdek and N. Pal. Cluster validation with generalized Dunn's indices. In *Proc. 2nd conf. on ANNES*, pages 190–193, Piscataway, NJ, 1995. IEEE Press.

[8] D. Blei and J. Lafferty. A correlated topic model of science. *The Annals of Applied Statistics*, 1(1):17–35, 2007.

[9] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[10] H. Chim and X. Deng. Efficient phrase-based document similarity for clustering. *IEEE Trans. on Knowl. and Data Eng.*, 20:1217–1229, September 2008.

[11] C. Cool and N. J. Belkin. A classification of interactions with information. In H. Bruce, R. Fidel, P. Ingwersen, and P. Vakkari, editors, *Emerging frameworks and methods. Proceedings of the Fourth International Conference on Conceptions of Library and Information Science (COLIS4)*, pages 1–15, Greenwood Village, 2002. Libraries Unlimited.

[12] D. Cutting, D. Karger, J. Pedersen, and J. Tukey. Scatter/gather: A cluster-based approach to browsing large document collections. In *Proc. SIGIR '92*, pages 318–329. ACM Press, 1992.

[13] H. Daumé III and D. Marcu. A Bayesian model for supervised clustering with the dirichlet process prior. *J. Mach. Learn. Res.*, 6:1551–1577, 2005.

[14] F. Diaz. Regularizing ad hoc retrieval scores. In *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 672–679, New York, NY, USA, 2005. ACM.

[15] A. El-Hamdouchi and P. Willett. Comparison of hierarchic agglomerative clustering methods for document retrieval. *The Computer Journal*, 22(3):220–227, 1989.

[16] D. Fudenberg and J. Tirole. *Game Theory*. MIT Press, 1983.

[17] N. Fuhr. A probability ranking principle for interactive information retrieval. *Information Retrieval*, 11(3):251–265, 2008. http://dx.doi.org/10.1007/s10791-008-9045-0.

[18] N. Fuhr and C. Buckley. Probabilistic document indexing from relevance feedback data. In *Proceedings of the 13th International Conference on Research and Development in Information Retrieval*, pages 45–61, New York, 1990.

[19] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proc. IJCAI'07*, pages 1606–1611, San Francisco, USA, 2007. Morgan Kaufmann Publishers Inc.

[20] G. Gordon. Hierarchical classification. In P. Arabie, L. Hubert, and G. Soete, editors, *Clustering and Classification*, pages 65–121. World Scientific, 1996.

[21] X. He, D. Cai, H. Liu, and W.-Y. Ma. Locality preserving indexing for document representation. In *Proc. SIGIR '04*, 2001.

[22] M. Hearst, A. Elliott, J. English, R. Sinha, K. Swearingen, and K.-P. Yee. Finding the flow in web site search. *Commun. ACM*, 45:42–49, September 2002.

[23] M. Hearst and J. Pedersen. Reexamining the cluster hypothesis: Scatter/gather on retrieval results. In *Proc. SIGIR '96*, pages 76–84. ACM Press, 1996.

[24] M. A. Hearst and E. Stoica. Nlp support for faceted navigation in scholarly collections. In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, pages 62–70, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

[25] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42:177–196, 2001.

[26] E. L. Ivie. *Search Procedures Based On Measures Of Relatedness Between Documents*. PhD thesis, Massachusetts Inst of Technology, 1966.

[27] D. M. Jackson. The construction of retrieval environments and pseudo-classifications based on external relevance. *Information Storage and Retrieval*, 6(2):187–219, 1970.

[28] N. Jardine and C. van Rijsbergen. The use of hierarchical clustering in information retrieval. *Information Storage and Retrieval*, 7(5):217–240, 1971.

[29] X. Ji and W. Xu. Document clustering with prior knowledge. In E. N. Efthimiadis, S. T. Dumais, D. Hawking, and K. Järvelin, editors, *SIGIR*, pages 405–412. ACM, 2006.

[30] M. Käki. Findex: search result categories help users when document ranking fails. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '05, pages 131–140, New York, NY, USA, 2005. ACM.

[31] S.-S. Kang. Keyword-based document clustering. In *Proc. IRAL '03*, pages 132–137, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

[32] E. Kanoulas, V. Pavlu, K. Dai, and J. A. Aslam. Modeling the score distributions of relevant and non-relevant documents. In *ICTIR '09*, pages 152–163. Springer, 2009.

[33] W. Ke, C. R. Sugimoto, and J. Mostafa. Dynamicity vs. effectiveness: studying online clustering for scatter/gather. In *Proc. SIGIR '09*, pages 19–26. ACM, 2009.

[34] J. Kleinberg. An impossibility theorem for clustering. In *Advances in Neural Information Processing Systems (NIPS)*, pages 446–453, 2002.

[35] O. Kurland. The opposite of smoothing: a language model approach to ranking query-specific document clusters. In *Proc. SIGIR '08*, pages 171–178. ACM, 2008.

[36] O. Kurland and C. Domshlak. A rank-aggregation approach to searching for optimal query-specific clusters. In *SIGIR '08*, pages 547–554, New York, NY, USA, 2008. ACM.

[37] O. Kurland and L. Lee. Respect my authority!: Hits without hyperlinks, utilizing cluster-based language models. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 83–90, New York, NY, USA, 2006. ACM.

[38] K. S. Lee, W. B. Croft, and J. Allan. A cluster-based resampling method for pseudo-relevance feedback. In *Proc. SIGIR'08*, pages 235–242. ACM, 2008.

[39] A. Leuski. Evaluating document clustering for interactive information retrieval. In *CIKM '01: Proceedings of the tenth international conference on Information and knowledge management*, pages 33–40, New York, NY, USA, 2001. ACM.

[40] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li. Rcv1: A new benchmark collection for text categorization research. *J. Mach. Learn. Res.*, 5:361–397, December 2004.

[41] W. Li, W.-K. Ng, Y. Liu, and K.-L. Ong. Enhancing the effectiveness of clustering with spectra analysis. *IEEE Trans. on Knowl. and Data Eng.*, 19(7):887–902, 2007.

[42] Y. Li, S. M. Chung, and J. D. Holt. Text document clustering based on frequent word meaning sequences. *Data Knowl. Eng.*, 64(1):381–404, 2008.

[43] X. Liu and W. Croft. Cluster-based retrieval using language models. In *Proc. SIGIR '04*, pages 186–193, New York, NY, USA, 2004. ACM.

[44] Y. Liu, W. Li, Y. Lin, and L. Jing. Spectral geometry for simultaneously clustering and ranking query search results. In *Proc. SIGIR '08*, pages 539–546. ACM, 2008.

[45] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[46] H. Nagamochi, T. Ono, and T. Ibaraki. Implementing an efficient minimum capacity cut algorithm. *Math. Program.*, 67(3):325–341, 1994.

[47] H. Nottelmann and N. Fuhr. From retrieval status values to probabilities of relevance for advanced IR applications. *Information Retrieval*, 6(4), 2003.

[48] T. Radecki. Mathematical model of time-effective information retrieval system based on the theory of fuzzy sets. 13:109–116, 1977.

[49] W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(366):846–850, 1971.

[50] S. E. Robertson. The probability ranking principle in IR. *Journal of Documentation*, 33:294–304, 1977.

[51] P. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, 20:53–65, November 1987.

[52] G. Salton, editor. *The SMART Retrieval System - Experiments in Automatic Document Processing*. Prentice Hall, Englewood, Cliffs, New Jersey, 1971.

[53] G. Salton and C. Buckley. Term weighting approaches in automatic text retrieval. 24(5):513–523, 1988.

[54] M. D. Smucker and J. Allan. A new measure of the cluster hypothesis. In L. Azzopardi, G. Kazai, S. E. Robertson, S. M. Rüger, M. Shokouhi, D. Song, and E. Yilmaz, editors, *ICTIR*, volume 5766 of *Lecture Notes in Computer Science*, pages 281–288. Springer, 2009.

[55] P. H. Sneath and R. R. Sokal. *Numerical Taxonomy*. Freeman, San Francisco, 1973.

[56] B. Stein and S. Meyer zu Eißen. Retrieval models for genre classification. *Scandinavian Journal of Information Systems (SJIS)*, 20(1):91–117, 2008.

[57] B. Stein, S. Meyer zu Eißen, and F. Wißbrock. On cluster validity and the information need of users. In *Proc. AIA 03*, pages 216–221. ACTA Press, 2003.

[58] A. Tombros and C. J. van Rijsbergen. Query-sensitive similarity measures for information retrieval. *Knowl. Inf. Syst.*, 6(5):617–642, 2004.

[59] A. Tombros, R. Villa, and C. V. Rijsbergen. The effectiveness of query-specific hierarchic clustering in information retrieval. *Information Processing and Management*, 38(4):559–582, 2002.

[60] C. J. van Rijsbergen. *Information Retrieval*. Butterworths, London, 2. edition, 1979.

[61] E. Voorhees. The cluster hypothesis revisited. In *Proc. SIGIR'85*, pages 188–196. ACM Press, 1985.

[62] M. Wu, M. Fuller, and R. Wilkinson. Using clustering and classification approaches in interactive retrieval. *Inf. Process. Manage.*, 37:459–484, May 2001.

[63] W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *Proc. SIGIR '03*, pages 267–273. ACM Press, 2003.

[64] K.-P. Yee, K. Swearingen, K. Li, and M. Hearst. Faceted metadata for image search and browsing. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '03, pages 401–408, New York, NY, USA, 2003. ACM.

[65] G. Yongming, C. Dehua, and L. Jiajin. Clustering XML documents by combining content and structure. In *ISISE '08*, pages 583–587. IEEE Computer Society, 2008.

[66] R. B. Zadeh and S. Ben-David. A uniqueness theorem for clustering. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, pages 639–646, Arlington, Virginia, United States, 2009. AUAI Press.

[67] O. Zamir and O. Etzioni. Grouper: A dynamic clustering interface to web search results. In *Proceedings of the 8th International Conference on Word Wide Web*, pages 1361–1374, 1999.

Appendix: Properties of the F_p measure

For demonstrating that F_p is a reasonable cluster quality metric, we show that it fulfills both the clustering metric axioms defined in [1] as well as the constraints defined in [3]. For that, let us first define document vectors denoting their relevance to the elements of the query set:

$$\vec{d_l}^T := (d_{l_1}, \ldots, d_{l_K}) \quad \text{with} \quad d_{l_i} = \begin{cases} 1, & \text{if } (q_i, d_l) \in \mathcal{R} \\ 0, & \text{otherwise} \end{cases}$$

With this definition, we can rewrite pairwise precision and recall as

$$P_p(D, Q, \mathcal{R}, \mathcal{C}) = \frac{1}{|D|} \sum_{\substack{C_i \in \mathcal{C} \\ c_i > 1}} \frac{1}{c_i - 1} \sum_{\substack{(d_l, d_m) \in C_i \times C_i \\ d_l \neq d_m}} \vec{d_l}^T \cdot \vec{d_m} \tag{11}$$

$$R_p(D, Q, \mathcal{R}, \mathcal{C}) = \frac{\sum_{C_i \in \mathcal{C}} \sum_{\substack{(d_l, d_m) \in C_i \times C_i \\ d_l \neq d_m}} \vec{d_l}^T \cdot \vec{d_m}}{\sum_{\substack{(d_l, d_m) \in D \times D \\ d_l \neq d_m}} \vec{d_l}^T \cdot \vec{d_m}} \tag{12}$$

As distance function w, we use the inverse of the scalar product of these document vectors:

$$w(d_l, d_m) = \frac{1}{\vec{d_l}^T \cdot \vec{d_m}} \tag{13}$$

In the following, we discuss each of Ackerman & Ben-David's axioms by first quoting their definition (adapted to our notation), and then demonstrating its validity for the pairwise F-measure.

Scale Invariance *Let D be a set of document vectors, called domain set in [1], and let w be a distance function. A quality measure m satisfies the scale invariance axiom if $m(\mathcal{C}, D, w) = m(\mathcal{C}, D, \lambda \cdot w)$ for all clusterings $\mathcal{C}$ of $\langle D, w \rangle$ and every positive λ.*

Due to the nature of OCF a scaling factor λ cannot be inserted directly as a parameter into F_p. To show that we nevertheless can satisfy this axiom we have to generalize the definition of the document vector components d_{l_i} to nonbinary, nonnegative relevance values (e.g. for reflecting degrees of relevance). Furthermore, we have to normalize precision by a global factor:

$$P'_p(D, Q, \mathcal{R}, \mathcal{C}) = \frac{P_p}{\sum_{\substack{(d_l, d_m) \in D \times D \\ d_l \neq d_m}} \vec{d_l}^T \cdot \vec{d_m}} \tag{14}$$

Then we can rescale distances by a factor of λ by multiplying all document vectors by its square root: $\vec{d'} = \sqrt{\lambda}\vec{d}$. It is easy to see that neither R_p nor P'_p are affected by rescaling.

Consistency *A quality measure m satisfies the consistency axiom if for every clustering $\mathcal{C}$ of $\langle D, w \rangle$ the fact that w' is a $\mathcal{C}$-consistent variant of w implies $m(\mathcal{C}, D, w') \geq m(\mathcal{C}, D, w)$. A distance function w' is called a $\mathcal{C}$-consistent variant of w, if $w'(x, y) \leq w(x, y)$ whenever x and y are in the same cluster in $\mathcal{C}$, and if $w(x, y) \geq w(x, y)$ otherwise.*

In our case, changing the distance between two documents is controlled by either reducing or increasing the number of queries both documents are relevant for. Obviously, a query set that leads to decreasing distances of documents that belong to the same cluster or increasing distances of documents that belong to different clusters can only improve P_p and R_p, and thus the clustering quality according to F_p.

Richness *A quality measure m satisfies the richness axiom if for every non-trivial clustering $\mathcal{C}$ of D, there exists a distance function w such that $\mathcal{C} = \text{argmax}_{\mathcal{C}' \in \mathbf{C}}(m(\mathcal{C}', D, w))$. A clustering of D is called trivial if it consists of either a single cluster or $|D|$ one-document clusters.*

We define a special query q_C per cluster $C \in \mathcal{C}$ such that all documents in C are relevant to q_C, while all documents in $D \setminus C$ are irrelevant. If the query set then is defined as $\{q_C \mid C \in \mathcal{C}\}$, the resulting document vectors entail a distance function w such that $m(\mathcal{C}, D, w)$ is maximum.

Isomorphism Invariance *A quality measure m is isomorphism-invariant if $m(\mathcal{C}, D, d) = m(\mathcal{C}', D, d)$ for all clusterings $\mathcal{C}$, $\mathcal{C}'$ of $\langle D, d\rangle$ that are isomorphic. Two clusterings $\mathcal{C}$ and $\mathcal{C}'$ are isomorphic if there exists a distance-preserving isomorphism $\varphi : D \to D$ such that $x, y \in C$ iff $x, y \in C'$, for all $x, y \in D$, $C \in \mathcal{C}$ and $C' \in \mathcal{C}'$.*

This axiom requires that clustering should be indifferent to the individual identity of the clustered elements—which holds since our quality metric does not refer to the identity of documents.

Now we regard the four constraints for cluster metrics as defined in [3], which can be regarded as preferences that should be satisfied by any reasonable cluster metric. We discuss each of the four constraints by first citing the definition from Amigo et al. (where the notation is adapted to ours), along with an example, and then testing it on our pairwise F-measure.

Constraint 1: Cluster Homogeneity *Let D be a set of items belonging to categories $L_1, \ldots, L_n$. Let $\mathcal{C}_1$ be a cluster distribution with one cluster C containing items from two categories L_i, L_j. Let $\mathcal{C}_2$ be a distribution identical to $\mathcal{C}_1$, except for the fact that the cluster C is split into two clusters containing the items with category L_i and the items with category L_j, respectively. Then an evaluation metric Q must satisfy $Q(\mathcal{C}_1) < Q(\mathcal{C}_2)$.* Example: $Q(aabb|\ldots) < Q(aa|bb|\ldots)$. (In our examples here, we assume that the '…' parts are equal for both distributions.)

Obviously, precision increases in this case, while recall remains the same, so this constraint is fulfilled.

Constraint 2: Cluster Completeness *Let $\mathcal{C}_1$ be a distribution such that two clusters C_1, C_2 only contain items belonging to the same category L. Let $\mathcal{C}_2$ be an identical distribution, except for the fact that C_1 and C_2 are merged into a single cluster. Then $\mathcal{C}_2$ is a better distribution:* $Q(\mathcal{C}_1) < Q(\mathcal{C}_2)$. Example: $Q(aa|aa|\ldots) < Q(aaaa|\ldots)$.

Here recall increases, while precision remains the same.

Constraint 3: Rag Bag *Let C_{clean} be a cluster with n items belonging to the same category. Let C_{noisy} be a cluster merging n items from unary categories (there exists just one sample for each category). Let $\mathcal{C}_1$ be a distribution with a new item from a new category merged with the highly clean cluster C_{clean}, and $\mathcal{C}_2$ another distribution with this new item merged with the highly noisy cluster C_{noisy}. Then $Q(\mathcal{C}_1) < Q(\mathcal{C}_2)$.* Example: $Q(aaaab|cdefg) < Q(aaaa|bcdefg)$.

Here precision is increased, while recall remains unchanged.

Constraint 4: Cluster Size vs. Quantity *Let us consider a distribution $\mathcal{C}$ containing a cluster C_l with $n+1$ items (for $n > 2$) belonging to the same category L, and n additional clusters $C_1 \ldots C_n$, each of them containing two items from the same category $L_1 \ldots L_n$. If $\mathcal{C}_1$ is a new distribution similar to $\mathcal{C}$ where each C_i is split in two unary clusters, and $\mathcal{C}_2$ is a distribution similar to $\mathcal{C}$, where C_l is split in one cluster of size n and one cluster of size 1, then $Q(\mathcal{C}_1) < Q(\mathcal{C}_2)$.* Example (with $n = 3$): $Q(aaaa|b|b|c|c|d|d) < Q(aaa|a|bb|cc|dd)$.

Here we would have $P_p(\mathcal{C}_1) = \frac{1}{3n+1}((n+1)\cdot 1 + n\cdot 2\cdot 0) < P_p(\mathcal{C}_2) = \frac{1}{3n+1}(n\cdot 1 + 1\cdot 0 + n\cdot 2\cdot 1)$. However, recall would decrease:

$$R_p(\mathcal{C}_1) = \frac{n(n+1)}{n(n+1)+n} \;>\; R_p(\mathcal{C}_2) = \frac{n(n-1)+n}{n(n+1)+n}.$$

Thus, we have to regard the ratio of the F_p values for this case: $F_p(\mathcal{C}_1)/F_p(\mathcal{C}_2) = (6n^2+13n+7)/(12n^2+9n)$. For $n > 1$, this ratio is smaller than 1, so we have $F_p(\mathcal{C}_1) < F_p(\mathcal{C}_2)$, as desired.