ABSTRACT
Entity annotation is emerging as a key enabling requirement for search based on deeper semantics: for example, a search on ‘John’s address’, that returns matches to all entities annotated as an address that co-occur with ‘John’. A dominant paradigm adopted by rule-based named entity annotators is to annotate a document at a time. The complexity of this approach varies linearly with the number of documents and the cost for annotating each document, which could be prohibitive for large document corpora. A recently proposed alternative paradigm for rule-based entity annotation [16], operates on the inverted index of a document collection and achieves an order of magnitude speed-up over the document-based counterpart. In addition the index based approach permits collection level optimization of the order of index operations required for the annotation process. It is this aspect that is explored in this paper. We develop a polynomial time algorithm that, based on estimated cost, can optimally select between different logically equivalent evaluation plans for a given rule. Additionally, we prove that this problem becomes NP-hard when the optimization has to be performed over multiple rules and provide effective heuristics for handling this case. Our empirical evaluations show a speed-up factor upto 2 over the baseline system without optimizations.

Categories and Subject Descriptors
H.4.m [Information Systems]: Miscellaneous

Keywords
Inverted index, evaluation plan, cost estimate

1. INTRODUCTION
Developments in semantic search technology [11, 14, 6] have motivated the need for efficient and scalable entity annotation techniques that can be deployed on a large scale. Entity annotation involves associating one of several well-defined labels with token sequences in unstructured documents. Example labels are ‘person name’, ‘organization’, ‘place’, ‘date’, etc. All rule based systems use a set of rules for the annotation purpose. A rule for entity annotation is a pair consisting of a pattern and a type. A pattern defines the sequence of tokens that need to be identified and the corresponding type denotes the type for the token sequence. In this paper, we only consider patterns that are regular expressions over tokens and the properties of tokens such as (i) which dictionaries they belong to, (ii) their part of speech and (iii) orthographic properties. Most current rule-based techniques [13, 1, 15] for this task operate at the document level, where each rule is evaluated against one document at a time. The computational complexity of this approach varies linearly with the number of documents and the cost for annotating each document, which could be prohibitive for large document corpora. Another drawback of this approach is that the entire document collection needs to be reprocessed for every minor modification made in the set of rules. An example modification is the addition of rules having common subexpressions with already evaluated rules.

Recently, [16] proposed a framework for annotating a document collection with typed entities by working solely on its inverted index. The framework yielded an order of magnitude speedup over a state-of-the-art document-based annotator [10]. The improvement achieved was even more pronounced when annotations had to be produced for a set of rules that were incrementally added. We will refer to this framework as IndexAnnot.

They use an inverted index that stores a posting list corresponding to each unique token in the document collection. A posting list contains all the position information for the token. They further define a set of operations on posting lists. The consint operation is one of the example operations used in IndexAnnot framework. The consint operation is a binary operation that takes two posting lists \( L_1 \) and \( L_2 \) corresponding to two tokens \( t_1 \) and \( t_2 \) as arguments. It returns a new posting list such that each entry points to a token sequence which consists of two consecutive tokens \( t_1 \) and \( t_2 \). The cost of a consint operation is dependent on the length and the nature of the posting lists given as its arguments.

The IndexAnnot framework translates the rules for annotation into a sequence of operations on the inverted index [23]. These sequence of operations are also known as evaluation plans. The associative and commutative property of index operations, define an equiva-
The focus of this paper is to develop methods for estimating the cost of the different evaluation plans and develop algorithms for finding the optimal plan. Based on the nature of index operations, we first define an equivalence class for an evaluation plan. Since, we need to determine the minimum cost evaluation plan without actually evaluating any evaluation plan, we also develop cost estimates techniques for the evaluation plans. We then present a polynomial time algorithm based on dynamic programming, that yields the minimum cost evaluation plan from the equivalence class. We further consider the problem of jointly optimizing the evaluation plans of a collection of rules, where the rules may have common sub-expressions. In this case, the overall optimization is complicated by the fact that the order in which the sub-parts are evaluated also affects the overall execution time. We prove that finding the optimal evaluation plan for this case is an NP hard problem, and develop effective heuristics for handling this situation.

The rest of this paper is organized as follows: in Section 2, we recap the basics of the index-based annotation approach. We also introduce the set of index operators required for evaluating the rules as well as define the concept of the AND/OR Tree for representing the evaluation plan. In Section 3, we formulate the optimization problem and present a dynamic programming algorithm for optimally ordering the nodes of the AND/OR Tree. We also present methods for estimating cost of each of the index operators used in the AND/OR Tree. In Section 4 we delve into the issues raised by multiple rules that share common sub-parts. We prove this is an NP hard problem and develop several heuristic algorithms for picking the best evaluation plan. Finally, in Section 5, we present our results that show an overall improvement up to a factor of 2 and Section 6 presents our conclusions.

2. BACKGROUND: ENTITY ANNOTATION USING INVERTED INDEX OPERATIONS

Figure 2 shows the process for entity annotation presented in [16]. A given document collection $D$ is tokenized and segmented into sentences. The tokens are stored in an inverted index $I$ [23]. The inverted index $I$ has an ordered list $I$ of the unique tokens $u_1, u_2, \ldots u_W$ that occur in the collection, where $W$ is the number of tokens in $I$. Additionally, for each unique token $u_i$, $I$ has a postings list.
Lu in this paper

Figure 2: Overview of the entity annotation process referred to in this paper

$L(u_i) \in L_1, L_2, \ldots, L_n$ of locations in $D$ at which $u_i$ occurs. $cnt(u_i)$ is the length of $L(u_i)$. Each entry $l_k$, in the postings list $L(u_i)$, has four fields: (1) a document identifier, (2) a sentence identifier, (3) the begin position of the particular occurrence of $u_i$, (4) the end position of the same occurrence of $u_i$. The entries in a posting list are sorted first by $d$, them by $id$, followed by $s$ and finally by $b$.

The input grammar used in [16] is the same as that used for named entity annotations in GATE [10]. The GATE architecture for text engineering uses the Java Annotations Pattern Engine (JAPE) [9] for its information extraction task. JAPE is a pattern matching language. [16] supports two classes of properties for tokens that are required by grammars such as JAPE: (1) orthographic properties such as an uppercase character followed by lower case characters, and (2) dictionaries (gazetteers) to which a token or a token sequence belongs. The instances of dictionary containment properties can be single tokens. However, instances of basic entities specified using orthographic properties must be single tokens. The different operation types are associated with gazetteer properties was developed in [16]. A rule for NE annotation may require a token to satisfy multiple properties such as Location Dictionary as well as Capsword. The posting list for tokens that satisfy multiple properties is created by performing an operation $\text{merge}(L_1, L_2, \ldots, L_n)$ which consists of two consecutive subsequences $\text{Opt}$ and $\text{Opt}$ within the same sentence, such that, $L_1$ has an entry for $\text{Opt}$ and $L_2$ has an entry for $\text{Opt}$. If either of $L_1$ or $L_2$ has its $\text{Opt}$ field set to true, the resultant posting list also has its $\text{Opt}$ field set to true. This operation is associative as well as commutative. Using these properties, the merge operation can be performed in a manner similar to the merge-sort algorithm. The cost of performing $\text{merge}(L_1, L_2, \ldots, L_n)$ operation is as follows:

$$\text{cost}_{\text{merge}} = O(\log(n) \sum_{i=1}^{n} |L_i|)$$

2.2 Operations on Postings List

Two operations on postings lists were defined in [16].

1. $\text{merge}(L_1, L_2, \ldots, L_n)$: Returns a posting list such that each entry in the returned list occurs in at least one of $L_1$ or $L_2$ ... or $L_n$. If any of $L_1$ or $L_2$ .... or $L_n$ has the $\text{Opt}$ field set to true, the resultant posting list also has its $\text{Opt}$ field set to true. This operation is associative as well as commutative.

2. $\text{consint}(L, L')$: Returns a postings list such that each entry in the returned list points to a token sequence which consists of two consecutive subsequences $\text{sa}$ and $\text{sb}$ within the same sentence, such that, $L$ has an entry for $\text{sa}$ and $L'$ has an entry for $\text{sb}$. If either of $L$ or $L'$ has its $\text{Opt}$ field set to true, the resultant posting list is merged with the corresponding optional posting list. If both $L$ and $L'$ have their $\text{Opt}$ fields set to true, the resultant posting list is merged with $L$ and $L'$ and the $\text{Opt}$ field of the result is set to true.

We will denote $\text{consint}(L, L')$ by $L \ast L'$.

There are several methods for computing $\text{consint}(L, L')$ depending on the relative size of $L$ and $L'$. If they are roughly equal in size, a simple linear pass through $L$ and $L'$ analogous to a merge, can be performed. If there is a significant difference in sizes, a more efficient modified binary search algorithm can be implemented. The details are shown in Figure 3.

Figure 3: The modified binary search algorithm for $\text{consint}$.
The Left-Deep AND/OR Tree is recursively built by scanning the regular expression from left to right and identifying every sub-regular expression for which a sub-tree can be built. Details of the algorithm that builds the Left-Deep AND/OR Tree from a regular expression are provided in [16].

Figure 4 shows an example regular expression and the corresponding Left-Deep AND/OR Tree; AND nodes are shown as circles.

\[
\text{cost} = \begin{cases} 
\min(\Lambda(L, L'), \Lambda(L', L), |L| + |L'|) & \text{if } L.\text{Opt} = f, L'.\text{Opt} = f \\
\min(\Lambda(L', L), |L| + |L'|) & \text{if } L.\text{Opt} = f, L'.\text{Opt} = t \\
\min(\Lambda(L, L'), |L| + |L'|) & \text{if } L.\text{Opt} = t, L'.\text{Opt} = f \\
(|L| + |L'|) & \text{if } L.\text{Opt} = t, L'.\text{Opt} = t 
\end{cases}
\]

where \( t \) and \( f \) stand for \text{true} and \text{false} respectively and \( \Lambda(X, Y) = 2|X/(\log_2(|Y|/|X|)) + 1 \).

2.3 Inverted Index-based Annotation using a Left-Deep AND/OR Tree

Each annotation pattern, which is a regular expression over basic entities is first translated into a Left-Deep AND/OR Tree [16]. A Left-Deep AND/OR Tree specifies a sequence of \text{consist} and \text{merge} operations on posting lists to obtain a postings list for the annotation. Associated with each node in the tree is a regular expression and a postings list that points to all the matches for the node’s regular expression in the document collection. There are two node types: AND nodes where the output list is computed from the consecutive intersection (\text{consist}) of the postings lists of two children nodes and OR nodes where the output list is formed by merging the posting lists of all the children nodes. Additionally, each node has two binary properties: Opt and selfLoop. The first property is set if the regular expression being matched is of the form ‘R?’, where '?' denotes that the regular expression R is optional. The second property is set if the regular expression is of the form ‘R+’, where '+' is the Kleen operator denoting one or more occurrences. For the case of ‘R+’, both properties are set.

We formally define an AND/OR Tree as follows:

\text{DEFINITION 1.} An AND/OR Tree is a directed tree \( T = (V, E) \) such that

1. Each vertex \( v \in V \) has the following attributes:
   - \text{type: Type of a vertex is one of the following: (1)AND, (2)OR and (3)LEAF.}
   - \text{Opt: A boolean indicating whether the subtree rooted at } v \text{ is optional.}
   - \text{selfLoop: A boolean indicating whether the subtree rooted at } v \text{ can repeat more than once.}
   - \text{signature: A string that denotes the regular expression associated with } v.

2. Every AND vertex \( v \) has exactly two children vertices.

If the right child of every AND node is a a leaf an OR node or an AND node with a self-loop, \( T \) is called a Left-Deep AND/OR Tree.

The Left-Deep AND/OR Tree is recursively built by scanning the regular expression from left to right and identifying every sub-regular expression for which a sub-tree can be built. Details of the algorithm that builds the Left-Deep AND/OR Tree from a regular expression are provided in [16].

Figure 4 shows an example regular expression and the corresponding Left-Deep AND/OR Tree; AND nodes are shown as circles.

Figure 4: An example regular expression and the corresponding Left-Deep AND/OR Tree

whereas OR nodes are shown as square boxes. Nodes having Opt and selfLoop properties are labeled with +, * or ?. The edges in the tree represent dependency between computing nodes. The main regular expression is at the root node of the tree. The leaf nodes correspond to basic entities and words.

Figure 5 outlines the algorithm for computing the postings list of a regular expression by operating bottom-up on the Left-Deep AND/OR Tree.

\begin{verbatim}
for Each node v in the reverse topological sorting of G_R do
  if v.nodetype == AND then
    Let v1 and v2 be the children of v
    L(v) = consist(L(v1), L(v2))
  else if v.type == OR then
    L(v) = merge(L(v.child1), ..., L(v.children))
  end if
  if v.selfLoop == 1 then
    L(v).Opt = 1
  end if
end for
\end{verbatim}

3. OPTIMIZING AND/OR Trees

3.1 Equivalence of AND/OR Trees

An AND/OR Tree defines a plan for computing the posting list for a regular expression on an inverted index \( I \). More specifically, the tree specifies a sequence of \text{consist} and \text{merge} operations that should be performed on posting lists using the algorithm outlined in Figure 5. The overall time for computing the posting list at the root is proportional to the number of accesses made to the posting lists involved.

\text{DEFINITION 2.} Given a sequence of operations \( S \) on posting lists from an index \( I \), the cost\( (S, I) \) denotes the number of accesses made to the posting lists while evaluating \( S \). For an AND/OR Tree \( G \), cost\( (G, I) \) is the cost of the sequence of operations performed on posting lists using the algorithm outlined in Figure 5.

We make the following important observations on the \text{consist} and \text{merge} operations:
1. The \textit{consint}(L, L') operator defined in Section 2.2 is associative i.e., \((L_1 \ast L_2) \ast L_3 = L_1 \ast (L_2 \ast L_3)\). However, \textit{consint} is not commutative. Thus, we cannot permute the order of \textit{consint} operations without changing the result. By virtue of the associative property of \textit{consint}, a sequence of \textit{consint} operations \(L_1 \ast L_2 \ast \ldots \ast L_k\) can be computed by parenthesizing the sequence in several equivalent ways. As an example, the parenthesization corresponding to a left-most evaluation first is \(((L_1 \ast L_2) \ast L_3) \ldots \ast L_k\). Each parenthesization of a chain of \textit{consint} operations leads to a different cost, depending on the sequence and the nature of intermediate posting lists.

2. The \textit{merge} operator is commutative and associative, i.e., \(\text{merge}(L_1, L_2, \ldots, L_n) = \text{merge}(\Pi(L_1, L_2, \ldots, L_n))\), where \(\Pi(L_1, L_2, \ldots, L_n)\) is any permutation of the posting lists. The number of possible permutations for a list of length \(n\) is \(n!\). The \textit{merge} operates on several posting lists simultaneously. The cost of evaluating \(\text{merge}(L_1, L_2, \ldots, L_n)\) is independent of the order of the posting lists being merged (c.f., Section 2.2).

The cost \((G, I)\) is sensitive to the ordering of \textit{consint} operations in \(G\) and does not depend on the ordering of the posting lists involved in the \textit{merge} operations. Based on this observation, we define a relation \(\Theta\) between two AND/OR Trees \(G_1\) and \(G_2\).

**Definition 3.** We say that \(G_1 \Theta G_2\) iff \(G_1\) and \(G_2\) are either the same or differ only in the parenthesization of \textit{consint} chains in \(G_1\) and \(G_2\).

The relation \(\Theta\) is symmetric, reflexive, and transitive by definition. Let us denote the equivalence class induced by \(\Theta\) on an AND/OR Tree \(G\) by \(\mathcal{G}_\Theta\). By virtue of the associativity property of the \textit{consint} operation, evaluation of any \(G' \in \mathcal{G}_\Theta\) will yield the same result. Figure 6 shows an example regular expression, the corresponding Left-Deep AND/OR Tree and an equivalent AND/OR Tree.

The value of \(\text{cost}(G, I)\) cannot be obtained without executing the algorithm in Figure 5. In practice, however, we would require to compare two plans for their costs before they are evaluated. We can estimate the value of \(\text{cost}(G, I)\) using cost estimates of \textit{consint} and \textit{merge} based on length of posting lists at leaves and length estimates of intermediate posting lists. Let \(\kappa(G, I, \rho)\) denote an estimated cost of evaluating \(G\) on \(I\), using some property \(\rho\). \(\rho\) is a function of \(G\) and \(I\) that estimates some property of the posting list that results from the evaluation of \(G\) on \(I\). As an example, \(\rho(G, I)\) could be the estimated length of the posting list that results from evaluating \(G\) on \(I\). \(\kappa(G, I, \rho)\) gives an estimate of the cost of evaluating \(G\) on \(I\), using the property function \(\rho\). We next define the problem of determining an optimal plan \(G_{\text{opt}} \in \mathcal{G}_\Theta\) for \(G\), given a cost estimation model \(\kappa\), a property \(\rho\) and the index \(I\). We refer to this problem as \(\text{OptPlan}\).

### 3.2 Problem Formulation

The plan specified by the Left-Deep AND/OR Tree \(G\) obtained using the algorithm outlined in [16] need not be optimal. There are many equivalent plans in the equivalence set \(\mathcal{G}_\Theta\) for \(G\), that could potentially have lower cost estimates than \(G\).

**Definition 4.** Let \(G\) be an AND/OR Tree and \(I\) an inverted index. Let \(\kappa(G, I, \rho)\) denote the cost of evaluating \(G\) on \(I\) using \(\kappa\) and a property \(\rho\) of \(G\) and \(I\). We define \(\text{OptPlan}\) as the problem of determining \(G_{\text{opt}} \in \mathcal{G}_\Theta\) that minimizes the estimated cost, \(\kappa(G_{\text{opt}}, I, \rho)\). That is, \(G_{\text{opt}} = \arg\min_{G \in \mathcal{G}_\Theta} \kappa(G, I, \rho)\).

Since \(\text{cost}(G, I)\) is only sensitive to the ordering of \textit{consint} operations in \(G\) and the length of intermediate posting lists, we will only consider cost estimates \(\kappa(G, I, \rho)\) that are determined by the \textit{consint} operations. We next define the Canonical-Flattened AND/OR Tree that represents every sequence of nodes operated by consecutive \textit{consint} as children of a common AND node. Thus, the Canonical-Flattened AND/OR Tree represents a \textit{consint} chain without the parentheses information.

**Definition 5.** A Canonical-Flattened AND/OR Tree is an AND/OR Tree that allows any number of children for every AND node. Every child of an AND node must have one of the following properties (1) \text{type is OR} (2) \text{type is LEAF} and (3) \text{type is AND and at least one of Opt or isSelfLoop options is set}. Figure 7 shows the Canonical-Flattened AND/OR Tree corresponding to the Left-Deep AND/OR Tree in Figure 6.

In Figure 8, we give the algorithm for converting any Left-Deep AND/OR Tree \(G_{LD}\) into the corresponding Canonical-Flattened AND/OR Tree \(G_{CF}\). The algorithm is invoked as \(\text{flattenTree}(G_{LD}, eList)\), where \(eList\) is an empty list. The algorithm does a depth first search traversal on \(G_{LD}\) and in the course, chains together every sequence of nodes that are operated on by consecutive \textit{consint} operations as sibling nodes of a common AND node in the resultant tree.
Note that a Canonical-Flattened AND/OR Tree could also be directly generated from the original regular expression.

![Algorithm](Figure 8: The algorithm for converting a Left-Deep AND/OR Tree $G_{LD}$ into the corresponding Canonical-Flattened AND/OR Tree $G_F$)

The OptPlan problem on $G_{LD}$ reduces to the following problem of OptConsintChain at each AND node of $G_F$.

**Definition 6.** Let $\Gamma = L_1 \ast L_2 \ast \ldots \ast L_n, n \geq 3$ be a chain of consint operations on posting lists and let $L_{1,n}$ denote the resultant posting list. Let $P_\Gamma$ be any parenthesization of $\Gamma$. Thus, $P_\Gamma$ determines a sequence in which consint is applied on the postings lists in $\Gamma$. Let $p$ be a property of $\Gamma$ such as the estimated length of the resultant posting list. The OptConsintChain problem is to determine a parenthesization $P_\Gamma$ of the list $\Gamma$ such that the cost estimate $k_{\text{consint}}(P_\Gamma, I, \rho)$ is minimized.

### 3.3 A dynamic programming algorithm

The number of parenthesizations for a sequence of $n$ consint operations is the Catalan number $[8] \frac{1}{n+1} \binom{2n}{n}$, which is exponential in $n$. We could go through each possible parenthesization (brute force), but this would require time $\Omega(4^n / n^{(n/2)})$, which is very slow and impractical for large $n$.

We can, however, apply dynamic programming to this problem. Let $\Gamma_{i,j}$ denote the consint chain between the $i^{th}$ and the $j^{th}$ posting lists. Let $\text{MinC}[i,j]$ be the minimum cost for computing $\Gamma_{i,j}$. The key observations that form the basis of the algorithm are:

1. The outermost parentheses partition the chain of consints between $i$ and $j$ at some $k, i \leq k < j$.

2. The optimal parenthesization order has optimal ordering on either side of $k$.

A recurrence for this problem is stated in Equation 3.3.

$$
\text{MinC}[i,j] = \min_{i \leq k < j-1} \{ \text{MinC}[i,k] + \text{MinC}[k+1,j] + \kappa_{\text{consint}}(G_{\text{consint}}^{\Gamma_{i,k}, \Gamma_{k+1,j}}, I, \rho) \}
$$

where $G_{\text{consint}}^{\Gamma_{i,k}, \Gamma_{k+1,j}}$ denotes a graph which has a consint node as a root node that has two children corresponding to $\Gamma_{i,k}$ and $\Gamma_{k+1,j}$. The function $\kappa_{\text{consint}}(\ldots)$ returns a cost estimate for the graph $G_{\text{consint}}^{\Gamma_{i,k}, \Gamma_{k+1,j}}$ using the estimates of property $\rho$ for the posting list $\Gamma_{i,k}$ and $\Gamma_{k+1,j}$. As updates are made to $\text{MinC}$, using values already computed, we need to note, that we need to determine the property values for these lists.

There are only $\binom{n}{2}$ substrings between 1 and $n$. Thus, it requires only $\Theta(n^2)$ space to store the optimal cost for each of them. We represent all the possibilities in a triangle matrix. We also store the value of $k$ in another triangle matrix $\text{MinSplit}$ to reconstruct the optimal parenthesization. The diagonal moves up to the right as the computation progresses. Figure 9 outlines our algorithm for computing the minimum cost parenthesization of the consint-chain $\Gamma$. The runtime for this algorithm is $O(n^3)$.

![Algorithm](Figure 9: The algorithm for computing the minimum cost parenthesization of a consint-chain $\Gamma$)

The method $\text{update}\_p(\rho[i,k], \rho[k+1,j])$ obtains a $\rho$ value for the posting list of $\Gamma_{i,j}$ based on the $\rho$ values for $\Gamma_{i,k}$ and $\Gamma_{k+1,j}$. Examples of $\rho$ will be discussed in Section 3.4. In particular, Equation 4 gives the possible update equations for $\text{update}\_p(\ldots)$.

#### 3.3.1 Algorithm for OptPlan

We now outline the algorithm that produces an AND/OR Tree $G_{\text{opt}}$ with the minimum cost estimate from a given Canonical-Flattened AND/OR Tree $G_F$. The algorithm, outlined in Figure 10, performs a depth-first traversal of $G_F$ and for every AND node $G$,
the \textit{consint}-chain formed by the children of $G$ is optimally parenthesized using the OptConsintChain algorithm (c.f. Figure 9). For every other node, OptPlan is recursively invoked.

\begin{align*}
\text{Procedure } \text{OptPlan} (G_F) \\
\text{Let } G \text{ be the root node of } G_F \\
\text{For Each } G' \in G.childList \text{ do} \\
\text{Call OptPlan on } G' \\
\text{end for} \\
\text{if } (G.type == \text{ AND}) \text{ then} \\
\text{Form a } \text{consint}-\text{chain } \Gamma \text{ comprising the children of } G \\
\text{Optimize } \Gamma \text{ using OptConsintChain to get an AND/OR Tree at } G \\
\text{end if}
\end{align*}

Figure 10: The algorithm for computing the minimum cost AND/OR Tree $G$ from a given Canonical-Flattened AND/OR Tree $G_F$.

3.4 Cost Estimates using Properties of Posting Lists

In this section, we define some cost estimation models. As mentioned in Section 3.1, we need to assess the cost of a plan even before it is evaluated. This is done using the property $\rho$ of the resultant list. In our implementation, we use the estimated length of the resultant list as the $\rho$. The $\rho$ estimate for the posting list $L_{\text{basic}}$ of any word or basic entity is the actual length of the posting list, i.e., $\rho(L_{\text{basic}}) = |L_{\text{basic}}|$. The $\rho$ estimate of the resultant list for a merge operation can be computed using the following equation:

$$\rho(G_{L_1, L_2, \ldots, L_n}, I) = \sum_{i=1}^{n} \rho(L_i)$$  \hspace{1cm} (3)

This approximation is an upper-bound on the length of the list resulting from a merge operation.

To determine the estimated length of the resultant list for a \textit{consint} operation, we use one of the following estimates of $\rho$ in the update function $\text{update}(\cdot)$:

$$\rho_{\text{min}}(G_{\text{consint}}, I) = \min(\rho(L_1), \rho(L_2))$$
$$\rho_{\text{am}}(G_{\text{consint}}, I) = \frac{\rho(L_1) + \rho(L_2)}{2}$$
$$\rho_{\text{gm}}(G_{\text{consint}}, I) = \sqrt{\rho(L_1) \times \rho(L_2)}$$
$$\rho_{\text{hm}}(G_{\text{consint}}, I) = \frac{2 \times \rho(L_1) \times \rho(L_2)}{\rho(L_1) + \rho(L_2)}$$  \hspace{1cm} (4)

Here, $\rho_{\text{min}}$, $\rho_{\text{am}}$, $\rho_{\text{gm}}$ and $\rho_{\text{hm}}$ are length estimates obtained by using the minimum, arithmetic mean, geometric mean and harmonic mean respectively. The $\rho_{\text{min}}$ estimate is not an upper-bound on the length of the resultant list from a \textit{consint} operation. This is because of the following reason. Given a regular expression, there can be multiple entries in its posting list that have the same \textit{first} or \textit{last} values. As an example, consider the posting list for $\text{CAPSWORD}^+$ which contains an entry for every consecutive sequence of capitalized words of length 1 or more; given a sequence of $n$ consecutive capitalized words in the document collection, there will be at least $n$ entries in the posting list that have the same \textit{start} value (similarly for \textit{end}). We therefore consider the other $\rho$ estimates to account for the length of both posting lists.

We obtain the cost estimate $\kappa_{\text{consint}}(G, I, \rho)$ based on length of posting lists at leaf nodes in $G$ and length estimates of intermediate posting lists. This is achieved by choosing the length estimates given in the above equations. The value of $\kappa_{\text{consint}}(G_{L_1, L_2}, I, \rho)$ is obtained using the following equation.

$$\kappa_{\text{consint}} = \begin{cases} 
\min(\Lambda(L, L'), \Lambda(|L| + |L'|) \text{ if } L.\text{Opt} = f, L'.\text{Opt} = f & \text{min} \\
\min(\Lambda(L, L'), \Lambda(|L| + |L'|) \text{ if } L.\text{Opt} = f, L'.\text{Opt} = t & \text{min} \\
\min(\Lambda(L, L'), \Lambda(|L| + |L'|) \text{ if } L.\text{Opt} = t, L'.\text{Opt} = f & \text{min} \\
\min(\Lambda(|L| + |L'|) \text{ if } L.\text{Opt} = t, L'.\text{Opt} = t & \text{min}
\end{cases}$$

where, $\Lambda(X, Y) = 2X(\log_2(\sqrt{Y/|X|} + 1))$.

4. OPTIMIZING MULTIPLE AND/OR Trees

Multiple AND/OR Trees might have common sub-sequences of evaluations. Figure 11 shows two Canonical-Flattened AND/OR Trees, $G_1$ and $G_2$ with nodes $R_2$ and $R_3$ that are effectively the same. Given this fact, $G_1$ and $G_2$ also share a sub-sequence of \textit{consint} operations, namely, $L(a) \ast L(R_2) \ast L(d)$. The application of OptPlan (c.f. Section 3.3.1) on $G_1$ or $G_2$ ignores this common sub-sequence information. In this section, we address the problem of optimizing multiple AND/OR Trees simultaneously.

\begin{figure}[h]
\centering
\includegraphics[width=0.4\textwidth]{figure11.png}
\caption{Two AND/OR Trees $G_1$ and $G_2$ having common sub-expressions.}
\end{figure}

In Section 3.2 we reduced the OptPlan problem to the OptConsintChain problem. Thus we can pose the problem of simultaneously optimizing multiple AND/OR Trees to the following problem of simultaneously optimizing multiple \textit{consint}-chains.

\textbf{Definition 7.} Let $S$ be a set, such that each $\Gamma \in S$ is a chain of \textit{consint} operations on posting lists and is of the form $\Gamma = L_1 \ast L_2 \ast \ldots \ast L_n, n \geq 3$. We define $MPlanOpt$ as the problem of determining the parenthesization $P_\Gamma$ of every $\Gamma \in S$ so that the cost estimate $\sum_{\Gamma \in S} \kappa(\Gamma, I, \rho)$ is minimized, where the cost for computing the posting list of any sub-expression $\beta$ of $P_\Gamma$ is considered only once.

4.1 NP-Hardness

In this section, we prove that the $MPlanOpt$ problem is NP-Hard.

\textbf{Theorem 1.} The $MPlanOpt$ problem is NP-Hard.

Instead of proving the Theorem 1, we prove the hardness result for the following simpler problem, RuleEvalWithCache.

\textbf{Definition 8.} Let $\Gamma = L_1 \ast L_2 \ast \ldots \ast L_n$ be a \textit{consint} chain. The problem of RuleEvalWithCache is to determine a parenthesization $P_\Gamma$ of the chain $\Gamma$ such that the cost estimate $\text{cost Est}(P_\Gamma, I)$ is minimized, where the cost for computing the posting list of any sub-expression $\beta$ of $P_\Gamma$ is considered only once.
The problem RuleEvalWithCache aims at evaluating a consint chain while caching the posting list of every intermediate sub-expression so that the cost of computing its posting list is accounted for only once. If any sub-expression is repeated in \( \Gamma \), the cached result of it can be re-used. As an example consider the consint chain \( \Gamma = C \ast D \ast X \ast Y \ast C \ast D \). The subexpression \( C \ast D \) occurs twice in the chain \( \Gamma \). The parenthesis operation \( ((C+D)\ast((X+Y)\ast(C+D))) \) of \( \Gamma \), is one of several possible parenthesizations that can gain from the caching of posting list of intermediate sub-expressions.

**Theorem 2.** The RuleEvalWithCache problem is NP hard.

**Proof.** We give a reduction from the smallest grammar problem (SmallGrammar), a well-known NP-Hard problem [7]. The SmallGrammar problem can be stated as follows: Given a string \( \sigma \), what is the smallest context free grammar that generates exactly \( \sigma \). The size of the grammar is defined to be the total number of symbols on the right sides of all production rules. For example, the smallest context free grammar generating the string: \( \sigma = \text{a rose is a rose is a rose} \) is as follows: \( 1) S \rightarrow BBA, (2) A \rightarrow \text{a rose} \) and \( 3) B \rightarrow A \).

There is a simple reduction from the SmallGrammar problem to RuleEvalWithCache. We omit details of the proof owing to space constraints.

### 4.2 Heuristics for solving MPlanOpt

Since MPlanOpt is an NP-hard problem, we resort to heuristics for solving this problem. All of our heuristics are based on one of the following two techniques:

1. **Caching during plan optimization:** Utilizing common subexpressions across AND/OR Trees during plan optimization.

2. **Caching during plan evaluation:** Storing in a hash map, the posting list for each evaluated node in an AND/OR Tree, using the node signature as a key.

Based on these two techniques, we developed following 3 heuristics.

1. **IOCache** : Order the AND/OR Trees randomly and optimize each plan individually. Evaluate rules in the same order and store in a hash map, the posting list of every evaluated node, using the corresponding node signature as the key. This heuristic uses only the second technique listed above.

2. **CCache** : Order the AND/OR Trees randomly and optimize plans in that order. After optimizing a plan using the OptConsintChain algorithm, the signature of every node in the optimized plan is cached. While optimizing subsequent plans, the cost estimate \( \kappa \) for a sub-expression \( \Gamma [i, j] \) is set to 0, if the sub-expression is already present in the cache.

3. **CSECache**: Store the signature of each node of every AND/OR Tree against its frequency. The cost estimate \( \kappa \) of any sub-expression is obtained by dividing the estimate obtained in Section 3.4 by the above frequency.

### 5. EXPERIMENTS

#### 5.1 Experimental setup

In this section, we present empirical comparison of performance of the optimized evaluation plans under various strategies, against the naive evaluation plan. The experiments were performed on three data sets, viz.

1. A combination of Reuters-21578 data set\(^1\) and the 20 News-groups data set\(^2\). The data set consists of 16331 non-blank documents after stripping off the header information. The size of this data set was 93 MB.

2. A subset of the TREC wt10G data set comprising of 167943 documents from directories WTX001 to WTX011 and from directories WTX025 to WTX033. We consider only the ‘title’ and ‘body’ fields of each document in our experiments. The size of this data set was 1 GB.

3. The enron email data set\(^3\). The data consists of 267978 documents after elimination of header information and discarding the resultant blank documents. The size of this data set was 1.2 GB.

Our entire implementation is in java. The experiments were performed on a dual 3.2GHz Xeon server with 4 GB RAM. The indexing was performed using Lucene [12]. Prior to indexing, the sentence segmentation and tokenization of each data set was performed using in-house Java versions of standard tools\(^4\).

Our rule specification is in JAPE [9]. JAPE is a version of CPSL (Common Pattern Specification Language). JAPE provides finite state transduction over annotations based on regular expressions. The JAPE grammar requires information from two main resources: (i) a tokenizer and (ii) a gazetteer.

We experimented with two manually crafted set of rules: (1) 12 rules for ‘Organization’ and (2) 15 rules for ‘Date’.

In the following subsection, we present results that we achieve by optimizing individual plans, while in Section 5.3, we present our results for the multi-plan optimization problem.

#### 5.2 Individual Plan Optimization

In our first experiment, we compare the performance of the individual plan optimization algorithm OptPlan (cf Section 3), abbreviated as IO, using different properties \( \rho \text{min}, \rho \text{gm}, \rho \text{hm} \) and \( \rho \text{am} \) against the baseline evaluation plan of the Left-Deep AND/OR Tree. The comparison is performed in terms of the actual number of accesses (cost) that are made to the posting list entries during all the consint operations that are performed at the time of evaluation. The experiments were performed for the ‘organization’ and ‘date’ rule sets.

Figure 12 compares the cost of OptPlan using the different properties against the baseline on the reuters+20NG, trec and the enron data sets. We observe that OptPlan performs consistently and substantially better than the baseline on all data sets, for both the rulesets and for all properties except \( \rho \text{am} \). The reason for the anomalous behavior of \( \rho \text{am} \) is as follows. The length of the posting list resulting from \text{consint}\((L_1, L_2)\) is generally more biased toward the length of the shorter of the two lists \( L_1 \) and \( L_2 \). The harmonic mean and geometric mean of \( |L_1| \) and \( |L_2| \) are also more biased.

\(^1\) http://www.daviddewitt.com/resources/testcollections/reuters21578/
\(^2\) http://people.csail.mit.edu/jrennie/20Newsgroups/
\(^3\) http://www.cs.cmu.edu/~enron/
\(^4\) http://l2r.cs.uiuc.edu/~cogcomp/tools.php
toward the length of the shorter of the two lists, while their minimum is determined only by the shorter list. Hence, these are good estimates of the length of the posting list resulting from consist. On the other hand, the arithmetic mean is equally biased by both the numbers and hence is not a good estimate of the length of the resultant list.

Figure 14 shows the performance gain we obtain on some individual rules for ‘organization’ and ‘date’ types on the reuters+20NG data set using $\rho_{\text{min}}$. The figure illustrates that the proposed algorithm is consistently better than the baseline across a variety of rules.

5.3 Multiple Plan Optimization
In our second experiment, we compare the performance of the heuristics that optimize multiple plans (cf. Section 4) against the baseline.

The measure used for comparison was the same cost as described above. For these experiments, we chose $\rho_{\text{min}}$ as the property, given its consistent and superior performance over the other properties.

Figure 13 compares the cost of the heuristics IOCache, COCache and CSECache against OptPlan and the baseline, on the reuters+20NG, trec and the enron data sets. We observe that IOCache shows a consistent speedup over OptPlan (which is always better than the baseline) and COCache performs consistently better than or as well as IOCache. However, this consistency is not observed in CSECache; this heuristic subsidizes the cost of computing sub-expressions common across plans, even before the optimal plans are computed and hence may not always perform well in practice.

6. CONCLUSION
In this paper we investigated the issues concerning optimization of evaluation plans using the inverted index. We developed a dynamic programming method that achieves an optimal evaluation plan for a
given rule in polynomial time. We further investigated the problem of multi-rule optimization and proved that the problem is NP hard. Several heuristic approaches for optimizing multiple rules were developed and implemented. Our experiments show a speed-up factor up to 2 over the baseline.

In our future work, we intend to develop optimization plans that also search in the equivalence class of evaluation plan including the distributive property of merge operation over consist operation. Better length estimates can be looked for. Including consist operation with gaps will further increase the evaluation plan search space, inviting the scope for even better results. Incorporation of such an operation can be made.

7. REFERENCES


