Friedrich-Alexander-Universität Erlangen-Nürnberg
Technische Fakultät, Department Informatik

ROLAND VALLERY
MASTER THESIS

TREE-BASED EDIT ANALYSIS OF WIKI ARTICLES

Submitted on 11 May 2016

Supervisors:
Dipl.-Inf. Hannes Dohrn
Prof. Dr. Dirk Riehle, M.B.A.
Professur für Open-Source-Software
Department Informatik, Technische Fakultät
Friedrich-Alexander-Universität Erlangen-Nürnberg

Versicherung

Ich versichere, dass ich die Arbeit ohne fremde Hilfe und ohne Benutzung anderer als der angegebenen Quellen angefertigt habe und dass die Arbeit in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegen hat und von dieser als Teil einer Prüfungsleistung angenommen wurde. Alle Ausführungen, die wörtlich oder sinngemäß übernommen wurden, sind als solche gekennzeichnet.

Erlangen, 11 May 2016

License

Erlangen, 11 May 2016

Abstract

With the aim to improve editing support, this thesis examines frequent patterns in the English Wikipedia's edit history. These patterns are assumed to resemble edit transformations Wikipedia authors had in mind when changing Wikipedia articles and might be of interest to facilitate text refactorings.

The transformations are expected to have a complex structure with multiple relations between elements of Wikipedia articles and edit operations, which cannot be trivially modeled. This thesis tackles this problem by encoding the information, which is hidden in the edit history, in graphs, representing edit operations and elements of Wikipedia articles as graph nodes and relations as links between these nodes. The resulting edit script graphs are mined for frequent subgraphs with the intention of retrieving interesting frequent patterns in the form of graphs.

As results we list, visualize and analyze the discovered frequent patterns and create pattern clusters, which resemble real-world text transformations at different levels of abstraction.

Contents

List of Figures

1 Introduction

Big data, allowing the extraction of new meaning from huge chunks of data, has gained widespread attention throughout recent years (Dean & Ghemawat, 2008; Abouzeid, Bajda-Pawlikowski, Abadi, Silberschatz, & Rasin, 2009). This thesis' contribution in this context is a structured analysis of the English Wikipedia's edit history.

In more detail, the main objective of this thesis is

- to convert the English Wikipedia's edit history into a sequence of edit scripts and
- to mine this sequence for recurring and interesting patterns of user behavior.

For this purpose this thesis covers the review of relevant literature as well as the actual crawling of the English Wikipedia's edit history based upon

- an existing tree-based format of wiki-content called Wiki Object Model (WOM) (Dohrn & Riehle, 2011) and
- the existing HD tree diff algorithm, which computes the difference between two documents and is especially suited for WOM based wiki articles (Dohrn & Riehle, 2014).

As resulting artifacts this thesis delivers

- a software that mines the edit script history for patterns and
- a summary of the discovered patterns.

Thereby, this thesis focuses on exploratory research and provides ground work for a detailed analysis of the found patterns for text refactorings.

2 Research Chapter

> Mechanical art's focus upon content—as opposed to form, aura and originality—is pertinent to digital art.
>
> — Melissa Langdon (Langdon, 2014)

2.1 Introduction

Since the launch of the English version of Wikipedia in 2001, more than 5 million articles have been written ("Wikipedia statistics", n.d.). Editing articles is tedious work, and it contains many repetitive tasks, such as formatting lists. Better interfaces and automation tools could potentially reduce effort and increase productivity (Dohrn & Riehle, 2013). However, this first requires solid understanding of typical editing processes in collaborative environments. One promising path towards this understanding is to observe and analyze past user behavior. A valuable source is Wikipedia itself, as besides the actual content, Wikipedia also stores the complete revision history of articles. This means the edit history contains the complete chronicle from a blank page to the current state of an article. In this huge pile of historical data, a vast amount of interesting knowledge could be hidden, which remains yet to be explored.

The challenge in analyzing these large amounts of material is to extract interesting information which is not blurred by the idiosyncrasies of the articles, but which represents recurring patterns that originate when transforming one revision into another.

This thesis mines the edit history of the English Wikipedia for frequent patterns of user behavior. It is based on the assumption that the edit history can be modeled as a graph structure and that frequent common subgraphs in this graph

structure represent recurring transformations the user had in mind on a higher abstraction level. In the course of this thesis we model a complete pattern mining process, consisting of

- applying the existing change detection algorithm HDDiff on Wikipedia article revision pairs and
- transforming the detected changes into graphs as well as
- mining these graphs for subgraphs and finally
- presenting the discovered frequent patterns

The structure of this thesis is as follows: first, related work and the state of the art in the field of pattern mining are reviewed in chapter 2.2.2. Based on this theoretical background the approach of using graphs to mine the edit history is justified in chapter 2.4. Then, a graph mining framework which retrieves patterns in the form of graphs is developed in chapter 2.5 and the actual mining is performed in chapter 2.6. The theses ends with a discussion of the mining results in chapter 2.7.

2.2 Related Work

2.2.1 State of the Art in Pattern Mining

The main goal of this thesis is to search the edit history of the English Wikipedia to extract unknown information about recurring transformations. This task falls under the scope of data mining, defined by Han (2005) as "the process of discovering interesting patterns and knowledge from large amounts of data". Han (2005) also describes the knowledge discovery (KDD) process which involves steps like data preprocessing, data mining, pattern evaluation and knowledge presentation. This thesis aims to model a complete instance of the KDD process for the purpose of mining frequent patterns in the English Wikipedia's edit history. Therefore, a careful look at existing techniques in the field of data mining is essential to make intelligent use of approaches which have already proven to be beneficial.

The basic approach to frequent pattern mining is the mining of frequent itemsets (Aggarwal & Han, 2014). A frequent itemset is a set of items which happen to occur frequently together as a group in a set of transactions which contain several items. However, the approach of mining frequent itemsets is only applicable to data in the form of a set and not for more complex data types, which fall under the category of dependency-oriented data as named by Aggarwal (2015). To mine complex data structures like graphs, as this thesis will accomplish, different mining algorithms have been developed. In particular, the problem of finding

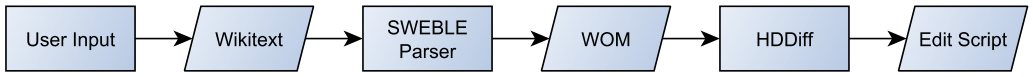


Figure 2.1: Edit script retrieval

frequent patterns in graphs can be tackled by algorithms which mine graphs for frequent subgraphs. Various algorithms for finding frequent subgraphs exist, for example gSpan (Yan & Han, 2002), GASTON (Nijssen & Kok, 2004) and SUBDUE (Ketkar, Holder, & Cook, 2005). However, the problem of finding common subgraphs is in general NP-hard (Gärtner, Flach, & Wrobel, 2003). Therefore, increasing the number of simultaneously mined graphs increases the run-time of the algorithms substantially. "ParSeMiS - the Parallel and Sequential Mining Suite" (n.d.) provides implementations of frequent subgraph mining algorithms.

Existing approaches to mine the English Wikipedia's edit history can be found in (Viégas, Wattenberg, & Dave, 2004), who visualize the activity on Wikipedia and focus on the procedural and collaborative side of Wikipedia, but do not explicitly search for edit transformations users had in mind.

The author is not aware of any existing attempts to mine the English Wikipedia's editing history by means of frequent subgraph mining, which constitutes the research approach of this thesis as laid out in section 2.4.

2.2.2 Wikitext Parsing and Edit Pattern Mining

While the last section gave a brief overview of the general state of the art in pattern mining, this section illustrates the context of this thesis and shows concrete existing techniques, which this thesis builds upon. Figure 2.1 shows how existing building blocks can be used to convert the English Wikipedia edit history into edit scripts, which forms the basis for the mining task this thesis performs. The first artifacts in this process are the revision texts, which Wikipedia authors have contributed to an article. These revisions are freely available in a format called wikitext.

The Sweble parser has been developed to convert wikitext into a tree-based object model called WOM (Dohrn & Riehle, 2011) The WOM allows to analyze the structure of articles at a higher abstraction level. To detect changes between

articles in the WOM format, the HDDiff comparison algorithm can be applied (Dohrn & Riehle, 2014). The HDDiff algorithm operates on revision pairs and creates an edit script, that enumerates the differences between the two revisions in form of a list of operations. If these operations are applied to the WOM of one revision, the other revision is retrieved. The HDDiff algorithm combines the advantages of tree-to-tree based comparison algorithms with the advantages of purely textual comparison algorithms by splitting text nodes to get a better matching between WOM nodes in the old and new revision. Thus, in contrast to common text diff algorithms, the edit scripts produced by HDDiff also feature move operations. The possible operations in the HDDiff edit script are insert, update, move and delete. In addition to the operations, the HDDiff edit script also refers to the nodes which are affected by the operations.

The edit scripts form the starting point of the pattern mining process performed in this thesis, which will be described in the following sections.

2.3 Research Question

This thesis performs exploratory research by retrieving possible frequent patterns from the English Wikipedia's edit history in a systematic and structured manner. The underlying research question is: *What are frequent edit operations performed by authors on Wikipedia?*

However, this thesis will not confirm the validity of the found patterns, meaning that the necessary confirmatory research of analyzing the relevance of the discovered frequent patterns is subject to future work. This confirmatory research could be performed by actual human beings manually assessing the true meaning of the mechanically found patterns.

2.4 Research Approach

This section describes the approach this thesis follows to retrieve frequent patterns of user behavior from the English Wikipedia's edit history. The reasoning behind our approach is based on a close look at the elements that change from one revision of an article to another. Analyzing these changes, multiple and complex associations between these elements can be noticed. For example, elements, which are inserted into the article to shape the new revision, can be linked to the elements surrounding the newly inserted elements, making up the context of the insertion.

Indo-European Family

Germanic Languages
- English
- Dutch
- German

Roman Languages
- French
- Spanish
- Basque

Figure 2.2: First revision of example article

In the course of this section we demonstrate how we cope with this complexity and how we strive to retrieve patterns from the English Wikipedia's edit history by using graph mining techniques. To this end we first explain the existing WOM format for Wikipedia articles. Then, we continue with a description of the existing change detection algorithm HDDiff and elucidate why we need to convert the outcome of the HDDiff algorithm applied on revision pairs into graphs for the purpose of retrieving interesting editing patterns. We conclude with a description of the methods we use to mine the constructed graphs for patterns.

As a use case, that is supposed to motivate and illustrate our research approach, the restructuring of a list in a made up article about language families is given in Figure 2.2. Throughout this chapter, the changes to this article are assumed to be exemplary transformations users perform on Wikipedia articles. The text in Figure 2.2 shall form an excerpt of the first version, i.e. the first revision, of this article.

2.4.1 The Wiki Object Model (WOM)

Wikipedia stores articles in a wiki markup language, which is used as input format by the authors, and which can be converted into HTML by the MediaWiki Software behind the English Wikipedia. In this wikitext format, the first revision of the example article can be formulated as shown in Figure 2.3.

To process the content of Wikipedia articles, a format which represents the elements of an article at a higher abstraction level than plain wikitext is advantageous. To this purpose the WOM (Wiki Object Model) format has been developed. The WOM format stores articles in a tree structure and provides means to operate on the article and its elements. The WOM can be represented in various

```
== Indo-European Family ==
Germanic Languages
* English
* Dutch
* German
Roman Languages
* French
* Spanish
* Basque
```

Figure 2.3: Wikitext of first revi sion

ways, for example, as an XML document or as a data structure in a programming language, and can be created from wikitext by the Sweble parser (see section 2.2.2). Figure 2.4 shows a graphical representation of the WOM of the example article, drawn with adapted existing visualization techniques. The meaning of the syntactical elements of the wikitext can be recognized in the WOM, e.g. the construct inside the markup "==" is presented as a heading. The symbol "\n" stands for a line break. Strings prefixing labels (for example "mww:") function as namespace identifiers. The colors used in the graphic will be explained later. The WOM also contains RTD (Round Trip Data) information, which preserve the original input of the user. For reasons of clarity, RTD elements are omitted in Figure 2.4.

2.4.2 The HDDiff Edit Script

As the research question states, this thesis' interest does not primarily lie in the Wikipedia articles themselves, but in the changes between article revisions. Therefore, this thesis applies the existing HDDiff change detection algorithm on pairs of article revisions in WOM format. The HDDiff algorithm matches elements of two revisions and determines the operations that are necessary to perform on the matched elements in order to convert one revision of the article into another. Figures 2.5 and 2.6 show the layout and the wikitext of a new revision of the example article and Figure 2.7 shows the resulting WOM tree of this revision created by the Sweble parser. Identical colors of nodes in Figures 2.4 and 2.7 indicate that the respective article elements have been matched by applying the HDDiff algorithm on these revision pairs. For a match elements don't necessarily have to be identical. Instead, it is sufficient if they are similar and form the best match. Non-leaf nodes are displayed in gray.

The actual result of the HDDiff algorithm is a so-called edit script. As explained

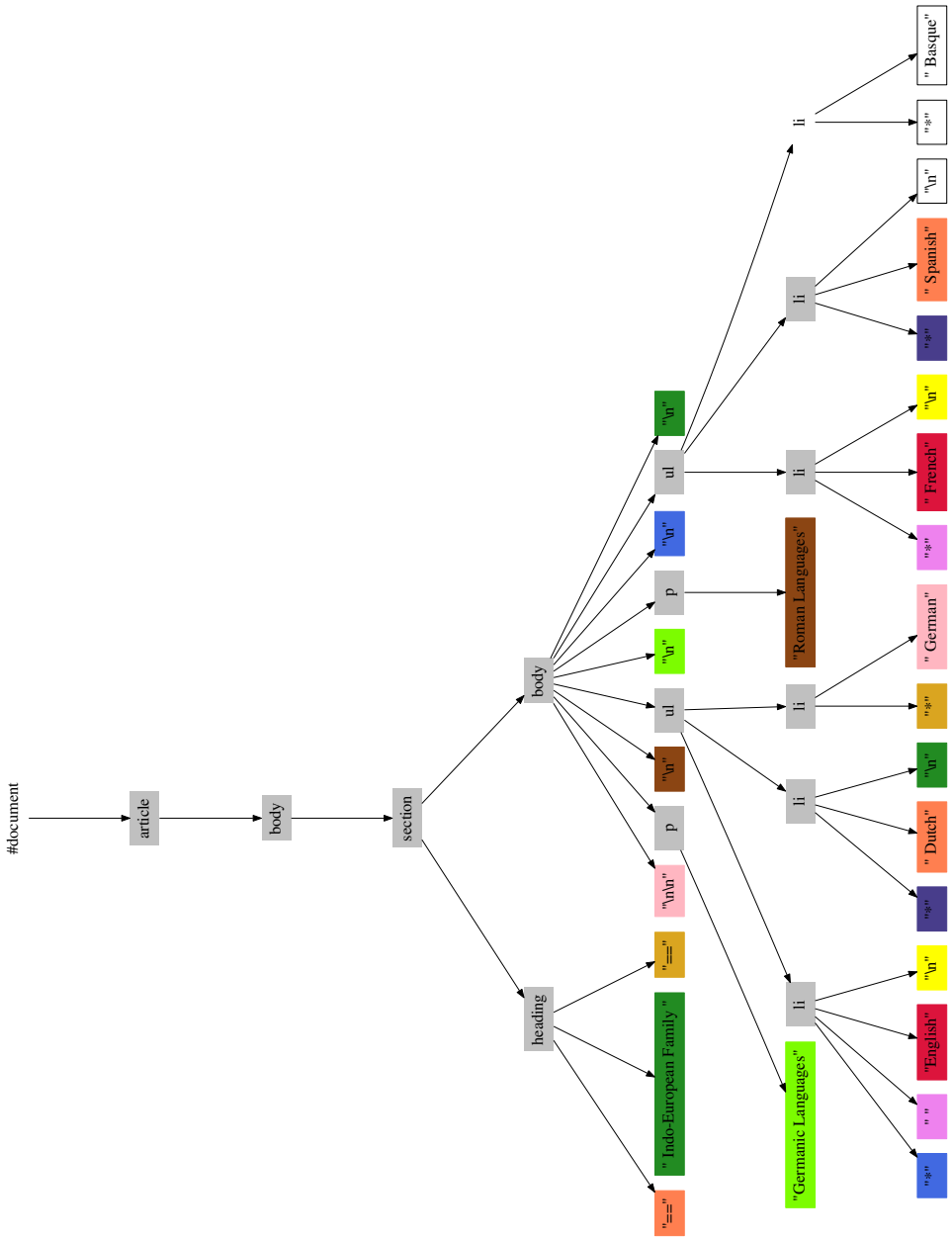


Figure 2.4: WOM of first revision

Indo-European Family

Germanic Languages
- English
- Dutch
- German

Roman Languages
- French
- Italian

Figure 2.5: Second revision of example article

```
== Indo-European Family ==
Germanic Languages
* [[English language—English]]
* Dutch
* German
Roman Languages
* French
* Italian
```

Figure 2.6: Wikitext of second revision

in section 2.2.2, an HDDiff edit script is a list of edit operations illustrating the changes from one document revision to another. Each operation in the edit script holds references to the nodes in the WOM that are affected by this operation. More concretely, the edit script for the two revision pairs of the example article is a list containing:

- 7 insert operations, one insert operation per inserted element. The inserted elements are part of an intlink and displayed in white color in Figure 2.7. An intlink is an internal link in a wiki software and is marked by double square brackets in the wikitext.
- 4 delete operations, one delete operation per deleted element. The deleted elements, belonging to the text "Basque", are displayed in white color in Figure 2.4.
- 1 move operation, as one element has been moved to a new location, i.e. the text node containing the text "English" has been made a child of the newly inserted target element of the intlink element.
- 1 update operation, as the substitution of the text "Spanish" by the text

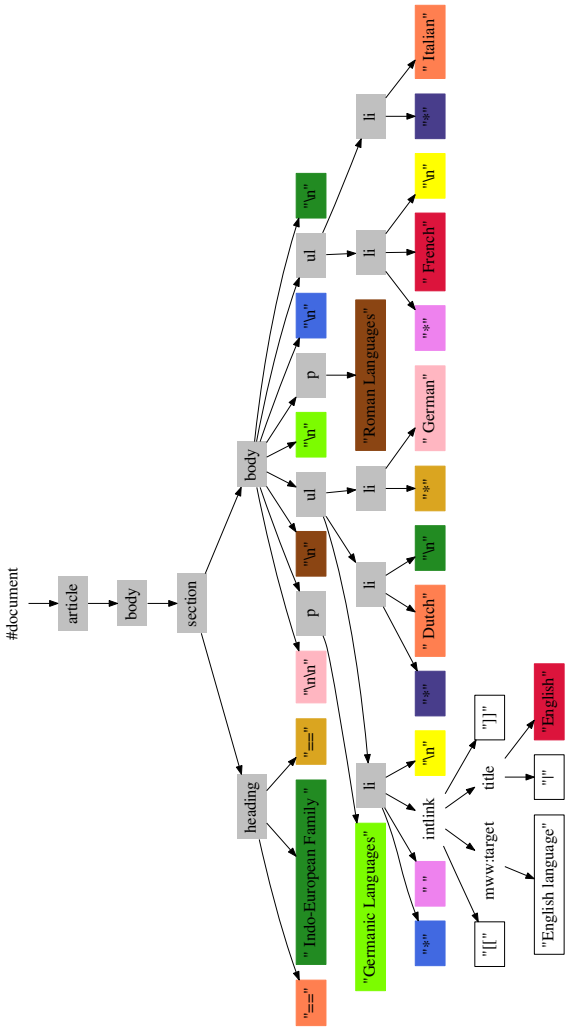


Figure 2.7: WOM of second revision

"Italian" has been reported as an update by the HDDiff algorithm.

2.4.3 Mining of Edit Script Graphs

The complex relations between article elements of different revisions are reflected in the HDDiff edit script, as described in the previous section. However, in order to apply common frequent pattern mining algorithms, the form of the data to be mined should be appropriate to serve as input for the mining algorithms. The edit script itself cannot be easily mined by conventional pattern mining algorithms, as the structure of the edit script is a plain list that does not give direct clues

about the frequent patterns to search for.

Therefore, we construct representations of the edit script that highlight the complex relations between WOM nodes and edit operations in the HDDiff edit script: *based on the reasonable assumption that graphs are naturally well suited to model structured data with multiple relationships (Aggarwal & Han, 2014), this thesis transforms HDDiff edit scripts into graphs and applies existing graph mining algorithms to discover frequent subgraphs in these graphs. The frequent subgraphs are supposed to correspond to operations Wikipedia authors had in mind on a higher abstraction level.*

2.4.4 Pattern Mining Workflow

Figure 2.8 gives an overview of the resulting work flow. We read random revision pairs from the English Wikipedia's edit history, call the HDDiff algorithm on these revision pairs and convert the resulting edit scripts into edit script graphs. Consecutively, the edit script graphs are passed to a frequent subgraph mining algorithm which determines the frequent subgraphs/patterns. Thus, the input to the subgraph mining algorithm consists of a set of graphs. Each of those graphs represents an edit script, created by applying the HDDiff algorithm on a random revision pair.

If a pattern occurs more than once in an edit script, the support count is incremented only by one for reasons of simplicity. For the same reason we restrict the found patterns to connected patterns.

The interesting subgraphs have to be frequent, meaning they must have a minimum support, i.e. they have to appear in a minimum percentage of edit script graphs.

Moreover, graphs with a larger number of vertices and edges are expected to represent more complex transformations. As trivial changes are not considered to be interesting this thesis only mines subgraphs with a minimum number of nodes.

Reasonable thresholds for the *minimum support* and the *minimum number of nodes* parameters have to be found during the mining process and will be given in section 2.6.

2.4.5 Pattern Clustering and Pattern Hierarchy

Some of the resulting patterns can be very similar, e.g. differing only in a single node. Therefore, and to reduce the complexity of a large number of patterns,

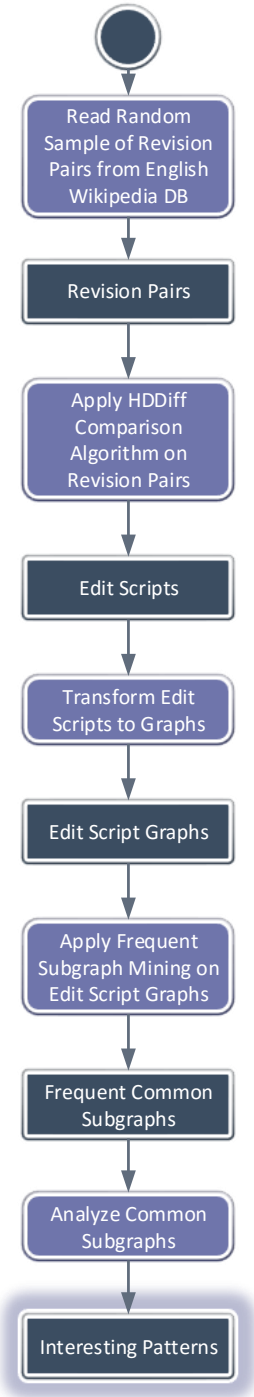


Figure 2.8: Pattern mining framework process

each pattern is mapped to a pattern cluster. Patterns with a similar structure (identified by similar subgraph relations) are assigned to the same cluster. Subsequently, the resulting clusters can be analyzed with respect to the subgraph relations between patterns to get a pattern cluster hierarchy. This hierarchy should demonstrate, to what extent larger transformations, represented by graphs with a larger number of vertices, consist of smaller transformations.

We propose to calculate the similarity of patterns based on the following distance function:

$$1 - \frac{\text{number of common common subpatterns and superpatterns of compared patterns}}{\text{number of all subpatterns and superpatterns of compared patterns}}$$

The two patterns to compare are themselves included in the sets of their respective subpatterns and superpatterns. Therefore, this function yields a value between 0 and 1. The smaller the return value, the more similar the patterns are.

2.5 Pattern Mining Framework

2.5.1 Purpose

In this chapter the pattern mining framework, which is a central contribution of this thesis, will be presented. This framework serves to mine and analyze frequent patterns in the form of graphs from the English Wikipedia's edit history. Furthermore, the derived conceptual model of the relevant entities (patterns, edit scripts, transformation algorithms, mining runs, etc.) constitutes a precise common terminology.

2.5.2 Requirements

The requirements for the pattern mining framework can be immediately derived from the research approach illustrated in section 2.4:

1. The framework has to implement the complete pattern mining process shown in Figure 2.8
2. The architecture of the framework should permit iterative improvements. This allows to repeatedly tune the configuration parameters which affect the pattern mining process.

3. The framework should offer different edit script to graph transformation algorithms, which allows specifying different characteristics of the patterns to be searched
4. The framework should include a graphical user interface to quickly assess the resulting patterns.
5. The framework should be smoothly integrated into the existing WOM and HDDiff infrastructure.
6. It would be advantageous if the framework provided support for subgraph relations among patterns to describe pattern hierarchies.

2.5.3 Architecture

The proposed framework follows a pipeline architecture, propagating data streams from pipeline step to pipeline step. This architecture enables the mining steps to save their results temporarily to file, thus avoiding constantly starting from the very beginning if a step fails to execute or if a step should be rerun with a different parameter configuration. The pipeline steps and the data structures, which are passed between the steps, correspond to the constituents of the mining process illustrated in Figure 2.8.

As the mining parameters have to be gradually adapted, the mining process is an iterative process, meaning that the steps can also be repeated from the very beginning, resulting in a loop structure. Another reason for the iterative nature of the process is that in a single mining run the number of input graphs that can be mined is rather limited. Due to computability restrictions the mining of an arbitrarily large number of graphs for frequent subgraphs is infeasible, especially because of the fact that the frequent subgraph mining problem is NP-hard (Gärtner et al., 2003).

2.5.4 Edit Script to Graph Transformation Algorithms

Before performing the actual mining, we have to convert the data to be mined, i.e. the edit scripts, into a form that is appropriate for the mining algorithms. This step is essential, as practical experience in the field of data mining suggests that data preprocessing accounts for a large part of the total effort of the overall mining process (Han, 2005). In our case the revision pairs, being converted into edit scripts by the HDDiff algorithm, form the basis for the mining process. As part of this thesis we propose two algorithms to transform edit scripts returned by the HDDiff algorithm into graphs, on which the frequent subgraph mining algorithms can be applied:

- the algorithm OPS_AS_NODES_BETWEEN_WOM_NODES, in which two WOM nodes are always connected by an edit operation in-between and
- the algorithm OPS_AS_LINKS_BETWEEN_TREES, which links the complete WOM tree of the old revision with the complete WOM tree of the new revision

The output of the transformation algorithms is restricted to the information present in the edit scripts. This means, for example, that the actual order of the edit operations the user carried through is not available, since the user is not permanently tracked while editing a page.

In the following sections we will describe the two developed transformation algorithms in detail exemplified by the fictional revision pair introduced in chapter 2.4. For both algorithms we only use labeled nodes and avoid labeling edges to reserve edge labels for future usages. Instead, we use labeled nodes functioning as role nodes between two nodes, thereby mimicking labeled edges between nodes. Furthermore, we only use undirected graphs to keep the structure of the graphs in a general format and support a broad range of mining algorithms.

The first algorithm is based on the idea to directly link each edit operation with the WOM nodes affected by this operation. Figure 2.9 shows the graph resulting from applying the OPS_AS_NODES_BETWEEN_WOM_NODES algorithm on the output of the HDDiff algorithm applied to the sample revision pair.

In Figure 2.9 the roles the WOM nodes play in an edit operation are drawn in green. These role nodes link the edit operation nodes, drawn in blue, with the WOM element nodes, drawn in red. The edit script graph can be divided into three connected components, one comprising the nodes involved in the insertion and moving of nodes and the other three components showing the deleted nodes, which are not connected to other nodes in the graph. The node for the intlink is inserted below its parent node - the "li" node - and serves as parent node for the target and title node of the link.

The second algorithm follows the instinct to connect matching nodes from the WOM trees of the compared revisions. Furthermore, the context of the nodes is added, i.e. the surrounding nodes in the WOM tree. This matching is specially designed to illustrate how move operations modify the WOM by shifting a WOM node to another destination in a different environment. The resulting edit script graph of the sample revisions can be seen in Figure 2.10. Red nodes indicate WOM nodes of the old revision and black nodes indicate WOM nodes of the new revision. The edit operations, drawn in blue, connect WOM nodes from the same or different WOM trees. The context of the nodes is marked by the green "parent" nodes.

In this section we have elucidated the purpose, design and output of the two

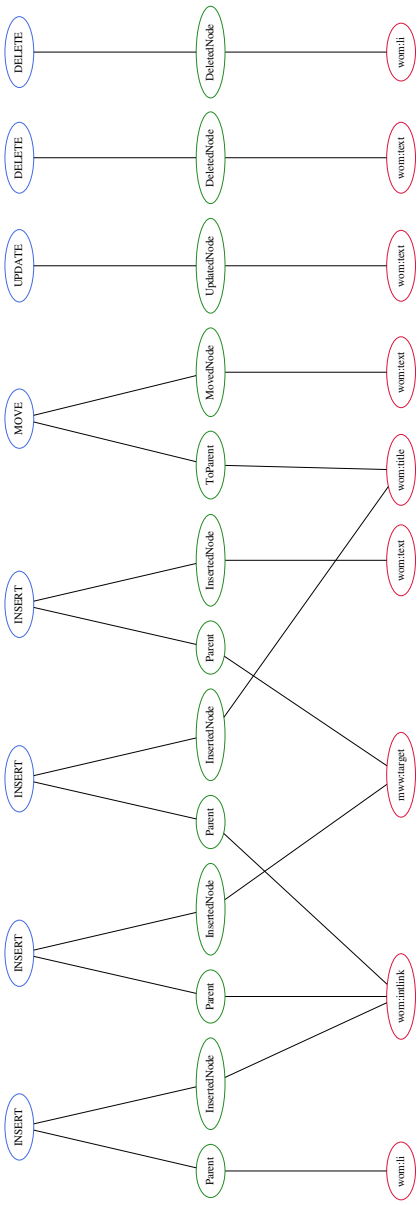


Figure 2.9: Edit script graph of edit operations as nodes algorithm

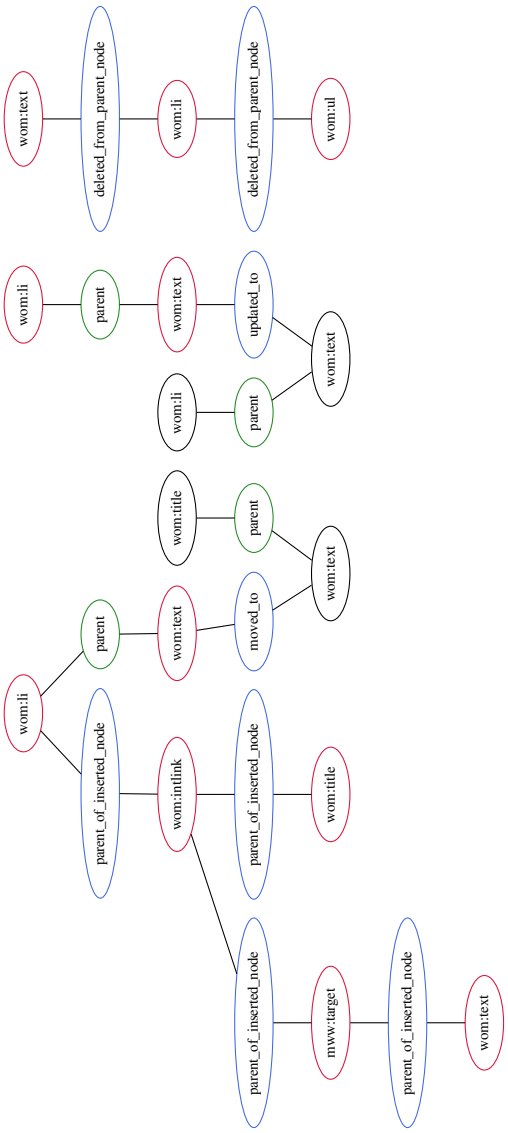


Figure 2.10: Edit script graph of edit operations linking WOM trees algorithm

developed edit script to graph transformation algorithms. Both algorithms have in common that they abstract away details of the actual textual content for the purpose of emphasizing commonalities in different articles. For example, the text string "English" does not appear in the edit script graphs and has been generalized into a wom:text node, which makes it possible for the mining algorithms to detect common structures in multiple graphs without being distracted by specific text strings.

2.5.5 Conceptual Model

Figure 2.11 shows the relevant classes of the model of the pattern mining framework and their relationships.

The loop structure of the mining framework is modeled by the **Run** entity, which contains the input configuration parameters of a single run of the mining loop, i.e. particularly *the minimum support* and the *minimum number of nodes* the frequent subgraphs have to fulfill. Results like the number of found patterns are also part of the Run class.

A set of **edit scripts**, which are arguably parts of runs, is the basic unit of input for the mining process. The minimum support parameter defines in how many edit scripts a pattern has to be found to be considered frequent. Edit scripts are uniquely identified by two revision IDs. A **graph transformation algorithm**, which is selected as a run parameter, converts an edit script into a graph. Thus, edit scripts as well as patterns are represented by **graphs**, which model complex relationships between article elements and edit operations.

Patterns, which are the main targets of this thesis, can be found in multiple runs, so that an entity **PatternRun** is created for each pattern found in a run. A pattern always belongs to a specific transformation algorithm, which is linked to the pattern via instances of the PatternRun and Run class. Patterns obtained by the same graph transformation algorithm can be assigned to a **cluster** with similar graphs.

As the same pattern can occur multiple times in the same edit script graph, these **PatternOccurrences** are instances of a PatternRun entity.

Subgraph relations among patterns are formed through a recursive relationship. If the graph of a pattern is a subgraph of an edit script graph, we regard the pattern itself as "occurring in the edit script", and if the graph of a pattern is a subgraph of the graph of another pattern, we regard the pattern itself as a "subpattern" of the other pattern, which is called the "superpattern".

Instances of all of the classes shown in Figure 2.11 - except the graphs belonging to edit scripts - are persisted to a database to keep track of the mining results,

allowing to verify the findings of this thesis. As edit script graphs can comprise a huge number of nodes and therefore claim huge amounts of storage, the edit script graphs are not stored to disk. Instead, the edit script graphs can easily be restored by running the HDDiff algorithm on the revision pairs again and subsequently reapplying the graph transformation algorithm on the returned edit scripts. The pattern graphs however are stored to disk, as these graphs form the mining results and have to be compared and matched to graphs gained from different runs. Patterns from different runs can be matched by reusing the frequent subgraph algorithm used in the mining process: the largest common subgraph of two patterns with the same number of nodes is the graph of the pattern itself if and only if the two patterns are identical.

The **Transformation** class, representing the actual transformations users had in mind, is drawn transparently in Figure 2.11 to demonstrate that this class is currently not part of the framework. The mapping from patterns to transformations is subject of future work and requires human intervention. Likewise, tags of a pattern will have to be added.

2.5.6 Implementation

The mining framework is implemented with standard Java EE technologies and includes a graphical web interface which visualizes the results and offers functionalities to the user to name and tag the discovered patterns. To mine the edit script graphs for frequent subgraphs we use the ParSeMiS (Parallel and Sequential Graph Mining Suite) framework ("ParSeMiS - the Parallel and Sequential Mining Suite", n.d.), which "search[es] for interesting substructures" and implements frequent subgraph mining algorithms. The source code of the ParSeMiS project is available online ("ParSeMiS project", n.d.).

2.5.7 Evaluation of the Pattern Mining Framework

In this section the developed pattern mining framework is evaluated against the requirements stated in chapter 2.5.2. In total, the developed pattern mining framework comprises the models of the relevant entities as well as the mining process, which are the key requirements. Thereby, the framework incorporates all steps of the KKD process outlined in chapter 2.2.1 for the purpose of mining the English Wikipedia's editing history. Furthermore, the pipeline architecture of the pattern mining framework allows to iteratively optimize the configuration parameters as requirement 2 postulated.

The design and implementation of two different edit script to graph transformation algorithms allows to compare and validate their results. The devel-

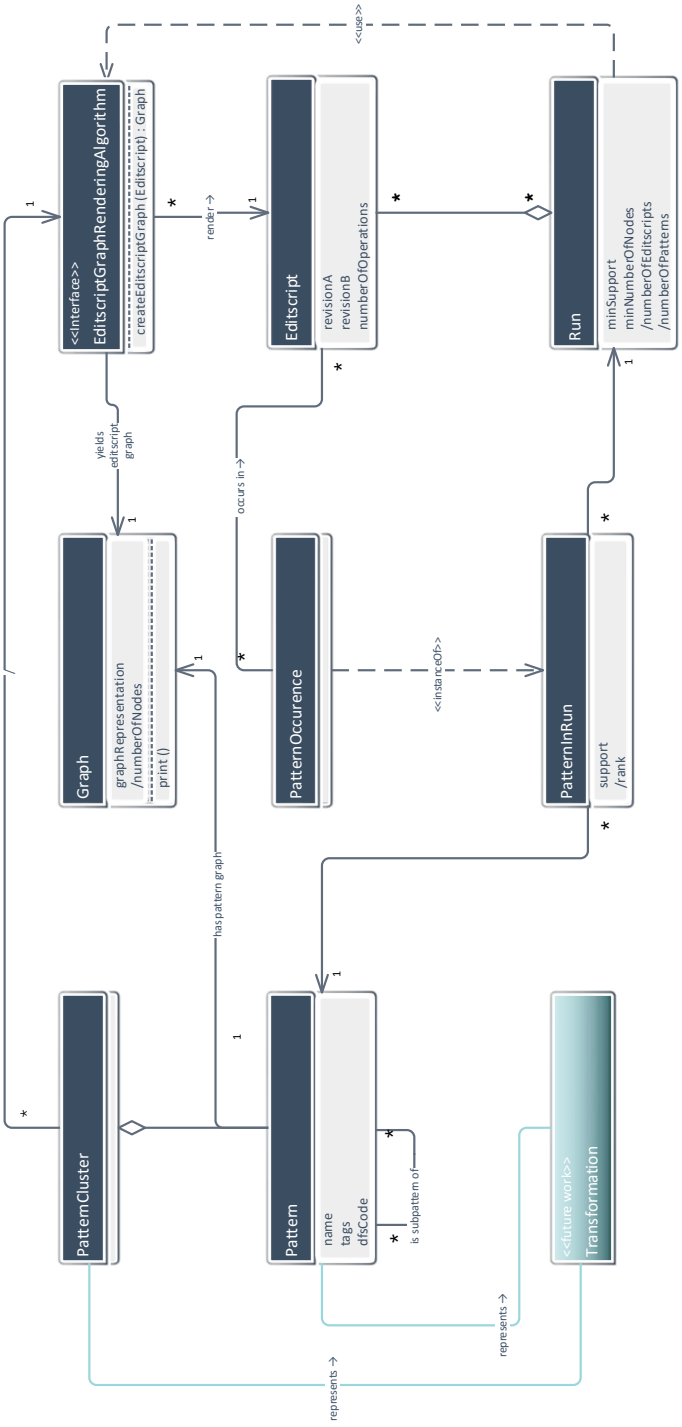


Figure 2.11: Pattern mining entities

oped framework includes a graphical user interface and connects existing components, particularly implementations of the WOM, the HDDiff algorithm and the ParSeMiS framework.

All in all the framework fulfills the requirements stated in section 2.5.2. Whether the framework is suited to mine frequent graph patterns from Wikipedia article revision histories will be evaluated in section 2.7.

2.6 Research Results

The following results are restricted to the graph algorithm OPS_AS_NODES_BETWEEN_WOM, as the alternative algorithm OPS_AS_LINKS_BETWEEN_TREES did not produce satisfying results, which will be discussed in the next section.

The mining process was decomposed into two phases:

1. In the first phase 134.936 edit script graphs were mined for frequent patterns in 323 runs with different mining parameters and 315 "pattern candidates" were gained in 2.819 edit script graphs. The revision pairs were restricted to immediately succeeding revisions of random articles. If ParSeMiS successfully mined the edit scripts and found frequent subgraphs satisfying the mining parameters of that run the frequent subgraphs were saved to file as pattern candidates. The highest number of edit script graphs ParSeMiS was able to process in a single run was 995. In this phase 257 runs of all 323 runs led to an out of memory error in ParSeMiS.

2. In the second phase, called the final summary run, the pattern candidates from phase one were validated. I.e. the pattern candidates were checked for occurrences in all of those edit script graphs which in phase one could be successfully mined for frequent subgraphs. This step was necessary to detect patterns that were only frequent by chance, i.e. because of the limited number of graphs in a single run. Phase two could be computed by checking pairs of pattern candidates and edit script graphs for a subgraph relation instead of checking all subgraph relations in a single run. In the final summary run 2.819 edit script graphs and 315 pattern candidates were checked for occurrences of the given pattern in the given edit script. Of those 887.985 possible combinations of edit script graphs and pattern candidates, 875.047 combinations (about 98%) could be successfully processed by ParSeMiS while 12.938 combinations led to an error when running ParSeMiS. 62.306 matches, i.e. occurrences of patterns in edit scripts, were found, not counting multiple occurrences of the same pattern in the same edit script graph.

The parameters used in the mining runs for the number of *edit script graphs* to mine, the *minimum node count* and the *minimum support*, can be found in Appendix C.

Figure 2.12 shows a scatter plot visualizing the relationship between the support percentage of a pattern in the final summary run and its number of vertices.

We checked all combinations of patterns for subgraph relations. Based on the discovered subgraph relations we generated the pattern hierarchy as proposed in section 2.4.5. The resulting pattern hierarchy is visualized in Figure 2.13. The arrows in this Figure indicate a part-of-relationship, e.g. at least one pattern in cluster 2 is a subpattern of at least one pattern in cluster 12 and so forth.

Detailed results including the largest patterns with a support count of at least 3 and the most frequent patterns can be found in the appendix.

2.7 Results Discussion

2.7.1 Evaluation

The first configuration parameter in the mining process, which was the edit script to graph transformation algorithm, showed unexpected results. Here, the graph transformation algorithm OPS_AS_LINKS_BETWEEN_TREES, which was designed to link the complete WOM tree of the old revision with the complete WOM tree of the new revision did not reveal meaningful insights. The patterns gained from this algorithm merely contained nodes of the WOM trees and mostly did not include the edit operations connecting the WOM tree nodes, i.e. the patterns only showed frequent article elements and missed the frequent transformations actually sought after. Therefore, this graph transformation algorithm was not pursued further. Instead, the focus was solely on the transformation algorithm OPS_AS_NODES_BETWEEN_WOM_NODES, in which two WOM nodes are always connected by an edit operation in-between.

An idiosyncrasy of the OPS_AS_NODES_BETWEEN_WOM_NODES algorithm is that some edit operations occupy more nodes in the edit script graph than others. For example, the update operation only refers to a single node (the updated node), whereas the insert operation refers to the inserted node as well as to the parent node of the inserted node, which leads to at least two nodes in the edit script graph instead of one. As a consequence of this idiosyncrasy, patterns with insert operations tend to be larger and are more likely to exceed the *minimum number of nodes* threshold. This makes the OPS_AS_NODES-_BETWEEN_WOM_NODES algorithm favor insert operations over update op-

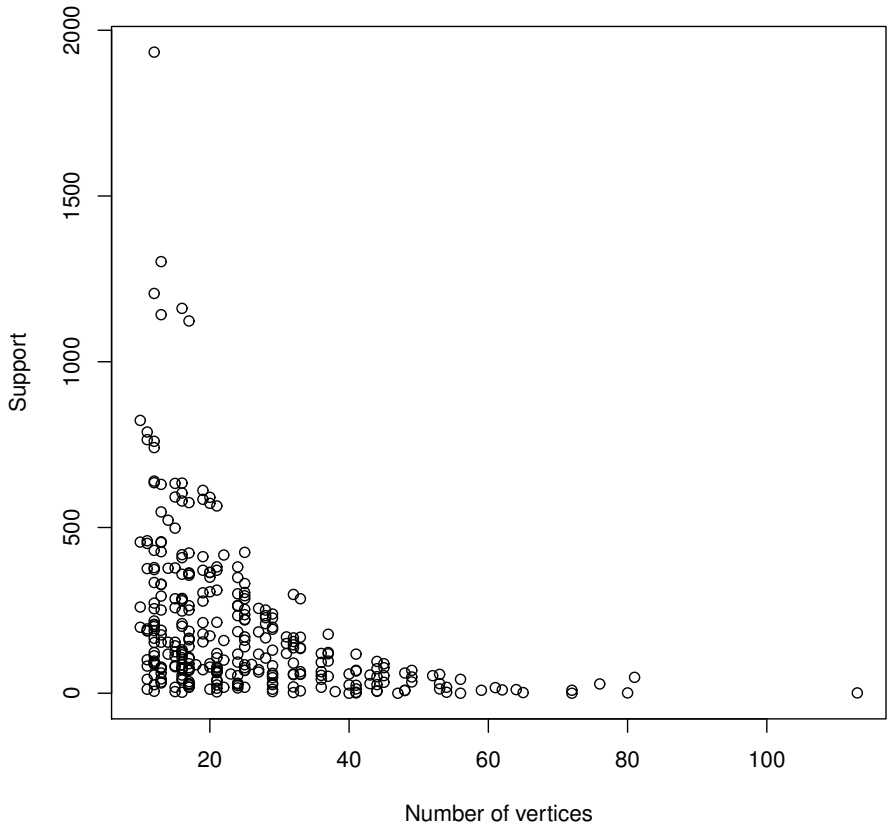
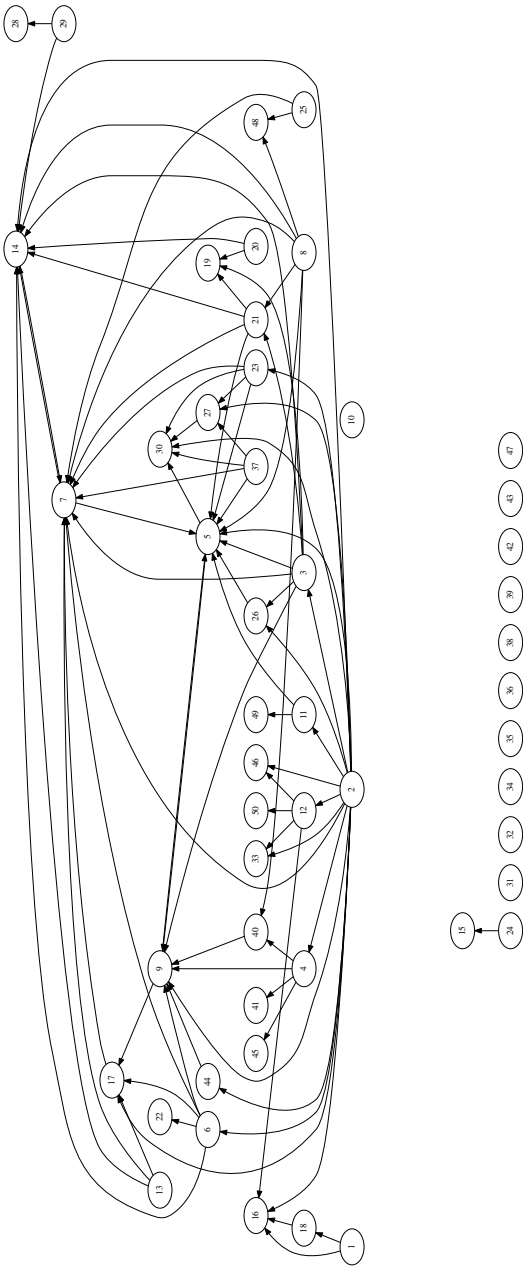


Figure 2.12: Pattern support and vertex number

Figure 2.13: Pattern cluster hierarchy



erations (regardless of this peculiarity, insert operations obviously occur much more frequently than all other operation types in the edit scripts).

Regarding the actual frequent subgraph mining step, the results showed that in 80% of the runs the mining process ended abruptly with an out of memory error. In consequence, the mining process had to be restricted to a set of 50 graphs in a single run, which is in fact not very high in magnitude. This indicates that the problem of finding common frequent subgraphs is in general too hard to allow mining a large number of edit script graphs simultaneously.

The patterns shown in the appendix cannot always be easily interpreted. In the current state of the pattern mining framework, it is non-trivial to recognize which actual transformation in the article refers to a found pattern. This is due to the fact that currently the nodes of a pattern cannot be exactly matched to the associated article elements, which would help the user in classifying the patterns. Instead, the user is left with the pattern graph and a list of revision pairs and their edit script graphs in which the pattern occurs but often it is unknown which text elements in the article are associated with which nodes in the pattern. Future versions of the developed graph mining framework might improve this shortcoming and keep track of the article elements associated with a node in pattern graphs and edit script graphs.

With respect to the clusters patterns were assigned to, it can be said that these clusters seem to be a good fit. At a first glance similar patterns appear to be assigned to the same cluster. The automatically generated pattern cluster hierarchy grants a first glimpse at the possible semantic meanings and relationships between the discovered patterns, indicating in how far larger transformations use smaller transformations as building blocks. However, a thorough evaluation of the proposed clustering remains future work.

2.7.2 Exemplary Discovered Patterns

In this section we present exemplary frequent patterns gained from the mining runs.

By far the most frequent pattern is pattern No. 6 (see Figure 2.15) which reflects the insertion of an internal link encompassing the target of the link. It does not come as a surprise that in a strongly connected document repository like Wikipedia, links make up the vast majority of discovered patterns (the number of occurences of pattern No. 6 is 1.934 in 2.819 revision pairs. The second frequent pattern, pattern No. 5, has a frequency of 1.302.)

The largest frequent pattern with a support count of above 3 is pattern number 186 (Figure 2.16), which stands for an interesting transformation as well: This

Figure 2.14: Transclusion

pattern shows the insertion of a transclusion with four attributes and values. A transclusion is the embedding of another document inside the current document and is often used for template mechanisms like infoboxes. Figure 2.14 shows an example infobox for a music band generated by a transclusion with attributes and values for origin, genre, Years active, Labels and Members.

2.8 Conclusion

In this thesis the problem of finding frequent patterns in the English Wikipedia's edit history was reduced to the problem of searching frequent subgraphs. Also, the relationships among these patterns were analyzed, particularly by exploiting subgraph relations. Future work can now focus on the visual inspection and manual or automated analysis of the discovered patterns.

The number of edit scripts successfully searched for frequent patterns - 2.819 - was quite small in comparison to the total number of articles on Wikipedia. This is mostly due to the research approach, which was based on graphs instead of sets. This approach has the advantage of giving the opportunity to model complex relationships, which this thesis exploited heavily by modeling structured relationships between article elements and edit operations. The downside of this approach however is that the resulting frequent subgraph mining problem is NP-hard, leading to a smaller number of edit scripts being mineable than by using sets or frequent itemsets.

Possible variants to the approach taken in this thesis are manifold. Future work could

- use different parameters, for example directed graphs instead of undirected graphs
- restrict edit script graphs to trees. That way, more efficient subgraph mining algorithms could be applied, for example SLEUTH (Zaki, 2005), which might discover rarer and larger common subgraphs as more edit script graphs could be included in a single run.
- use data types completely different from graphs. These types could be sets for frequent itemset mining or completely unstructured data which could for example be processed by the commercial IBM 'Watson' software (Ferrucci, 2011).
- analyze the whole edit history of a single article or even between articles instead of only considering changes between two revisions of the same article

Future work might also take more semantic aspects into account. Whereas this thesis focused on purely structural changes (e.g. the indentation of a list), a different promising approach might be to shed more light on the semantic content that has changed (e.g. why the list was indented).

Appendix A Table of Most Frequent Patterns

Id	Vertex Count	Support	Cluster	Transformation Name
6	12	1934	2	insertion of intlink with target with text
5	13	1302	2	insertion of intlink with target attribute below paragraph
8	12	1206	2	insertion of intlink with target and other item below paragrah
4	16	1161	2	insertion of intlink with target and other item below new paragraph
7	13	1142	2	insertion of intlink with target
3	17	1123	2	insertion of intlink with target and text below new paragraph
9	10	823	2	insertion of three items below paragraph
15	11	788	2	insertion of text and two other elements below paragraph
19	11	765	3	insertion of an element and a paragraph with child element below body
16	12	760	3	insertion of text and paragraph with child element below body
13	12	741	2	insertion of two text nodes and another element below paragraph
17	12	640	2	insertion of intlink with title and target below parent element
66	12	635	3	insertion of element and paragraph with text below body
12	16	634	2	insertion of intlink with title and target with text below element
18	15	633	2	insertion of link with target and two other elements below paragraph
26	13	630	3	insertion of text and paragraph with text below body
11	19	612	2	insertion of intlink with target with text and text element 2 other elements below paragraph

Id	Vertex Count	Support	Cluster	Transformation Name
14	16	604	2	insertion of text and intlink with target and another element below paragraph
23	15	592	2	insertion of intlink with title and target below element
10	20	591	2	insertion of intlink with target with text and text element and other element below paragraph
21	19	585	2	insertion of intlink with title and target below other element
22	16	580	2	insertion of intlink with title with text and target below other element
69	17	575	2	insertion of 2 text elements and an intlink with target below paragraph
20	20	573	2	insertion of intlink with title with text and target with text below other element
44	21	565	2	insertion of 2 text elements and an intlink with target with text below paragraph

Appendix B Table of Largest Frequent Patterns

Id	Vertex Count	Support	Cluster	Transformation Name
186	81	48	18	insertion of transclusion with 4 attributes with name and value
78	76	28	16	insertion of transclusion with two argument with name and value and one argument with value
187	72	9	16	insertion of named transclusion with 5 arguments with value
299	64	11	9	insertion of definition list item with 2 intlink2 with title and target and 4 text elements and another element below definition list

Id	Vertex Count	Support	Cluster	Transformation Name
278	62	10	5	insertion of text and 3 intlinks with target and title and 2 other elements below paragraph
117	61	17	9	insertion of definition list with two list items of which one has 3 text elements and 2 intlinks with title and target
143	59	9	5	insertion of a paragraph with 2 intlinks with target and 4 text elements and another element below body
121	56	42	5	insertion of 2 intlinks with target and title and 3 text elements and another element below paragraph
144	54	18	5	insertion of an intlink with target and title and another intlink with title and 4 text elements and another element below paragraph
246	54	3	9	insertion of definition list with list item with bold text and 4 other text elements and another element
149	53	29	5	insertion of 4 text elements and 3 intlinks with target below paragraph
235	53	13	17	insertion of list item with intlink with target and 3 text nodes below unordered list
87	53	57	5	insertion of 3 intlinks with target and title and 2 text elements below paragraph
145	52	53	5	insertion of section with heading and body with inserted text and inserted paragraph with 2 text elements and intlink with target and title below body
220	49	34	5	insertion of paragraph with 3 intlinks with target and title and a text element below paragraph

Id	Vertex Count	Support	Cluster	Transformation Name
88	49	49	5	insertion of 3 intlinks with target and 3 text elements below paragraph
188	49	69	18	insertion of transclusion with 4 arguments with value
79	48	61	5	insertion of an element and 2 intlinks with target and an intlink with title and target below paragraph
164	48	9	9	insertion of definition list with a definition list item with 3 text elements and an intlink with target and title and another element and a text element below body
175	48	8	5	insertion of a paragraph with text and 2 other text elements and a paragraph with an intlink with target and with 3 text elements and another element below body
267	45	89	5	insertion of 3 intlinks with target and two text elements below paragraph
123	45	33	17	insertion of a list item below with 3 intlinks with target below unordered list
185	45	53	14	insertion of section with heading with text and body with 2 text elements and an unordered list with a list item with an intlink with target below body
119	45	77	5	insertion of section with heading and body with text and paragraph with intlink with target and two text elements below body
297	44	74	5	insertion of paragraph with 2 text elements and an intlink with target and an intlink with target and title below other element

Appendix C Table of Runs

Id	Number Of Found Patterns	Number Of Edit Scripts	Minimum Node Count	Minimum Support	Exception Occured
Summary	315	2683			
322	0	50	10	12	TRUE
321	0	50	10	12	TRUE
320	0	50	10	12	TRUE
319	0	50	10	12	TRUE
318	0	50	10	12	TRUE
317	0	50	10	12	TRUE
316	0	50	10	12	TRUE
315	0	50	10	12	TRUE
314	24	48	10	4	FALSE
313	0	50	10	4	TRUE
312	24	47	10	4	FALSE
311	0	50	10	4	TRUE
310	0	50	10	4	TRUE
309	0	50	10	4	TRUE
308	0	50	10	4	TRUE
307	0	50	10	4	TRUE
306	0	50	10	4	TRUE
305	0	50	10	12	TRUE
304	0	50	10	12	TRUE
303	0	50	10	12	TRUE
302	17	47	10	4	FALSE
301	0	50	10	4	TRUE
300	0	50	10	4	TRUE
299	0	50	10	4	TRUE
298	18	47	10	4	FALSE
297	0	50	10	4	TRUE
296	0	50	10	4	TRUE
295	0	50	10	4	TRUE
294	0	50	10	4	TRUE
293	32	47	10	4	FALSE
292	0	50	10	4	TRUE
291	26	48	10	4	FALSE
290	0	50	10	4	TRUE
289	0	50	10	4	TRUE
288	0	50	10	4	TRUE

Id	Number Of Found Patterns	Number Of Edit Scripts	Minimum Node Count	Minimum Support	Exception Occured
287	0	50	10	4	TRUE
286	0	50	10	4	TRUE
285	14	47	10	4	FALSE
284	0	50	10	4	TRUE
283	0	50	10	4	TRUE
282	0	50	10	4	TRUE
281	0	50	10	4	TRUE
280	0	50	10	4	TRUE
279	0	50	10	4	TRUE
278	0	50	10	4	TRUE
277	0	50	10	4	TRUE
276	0	50	10	4	TRUE
275	0	50	10	4	TRUE
274	0	50	10	4	TRUE
273	0	50	10	4	TRUE
272	0	50	10	4	TRUE
271	0	50	10	4	TRUE
270	0	50	10	4	TRUE
269	0	50	10	4	TRUE
268	48	46	10	4	FALSE
267	0	50	10	4	TRUE
266	0	50	10	4	TRUE
265	0	50	10	4	TRUE
264	0	50	10	4	TRUE
263	0	50	10	4	TRUE
262	0	50	10	4	TRUE
261	0	50	10	4	TRUE
260	0	50	10	4	TRUE
259	0	50	10	4	TRUE
258	0	50	10	4	TRUE
257	0	50	10	4	TRUE
256	0	50	10	4	TRUE
255	0	50	10	4	TRUE
254	0	50	10	4	TRUE
253	0	50	10	4	TRUE
252	20	48	10	4	FALSE
251	0	50	10	4	TRUE
250	0	50	10	4	TRUE
249	0	50	10	4	TRUE

Id	Number Of Found Patterns	Number Of Edit Scripts	Minimum Node Count	Minimum Support	Exception Occured
248	22	47	10	4	FALSE
247	0	50	10	4	TRUE
246	0	50	10	4	TRUE
245	17	47	10	4	FALSE
244	19	48	10	4	FALSE
243	22	47	10	4	FALSE
242	20	47	10	4	FALSE
241	0	50	10	4	TRUE
240	46	48	10	4	FALSE
239	0	50	10	4	TRUE
238	0	50	10	4	TRUE
237	0	50	10	4	TRUE
236	0	50	10	4	TRUE
235	0	50	10	4	TRUE
234	0	50	10	4	TRUE
233	0	50	10	4	TRUE
232	0	50	10	4	TRUE
231	0	50	10	12	TRUE
230	13	48	10	4	FALSE
229	0	50	10	4	TRUE
228	0	50	10	4	TRUE
227	0	50	10	4	TRUE
226	0	50	10	4	TRUE
225	16	49	10	4	FALSE
224	0	50	10	4	TRUE
223	0	50	10	4	TRUE
222	0	50	10	4	TRUE
221	0	50	10	4	TRUE
220	0	50	10	4	TRUE
219	0	50	10	4	TRUE
218	0	50	10	4	TRUE
217	0	50	10	4	TRUE
216	37	45	10	4	FALSE
215	0	50	10	4	TRUE
214	0	50	10	4	TRUE
213	0	50	10	4	TRUE
212	0	50	10	4	TRUE
211	20	48	10	4	FALSE
210	0	50	10	4	TRUE

Id	Number Of Found Patterns	Number Of Edit Scripts	Minimum Node Count	Minimum Support	Exception Occured
209	0	50	10	4	TRUE
208	0	50	10	4	TRUE
207	0	50	10	4	TRUE
206	0	50	10	4	TRUE
205	0	50	10	4	TRUE
204	28	44	10	4	FALSE
203	0	50	10	4	TRUE
202	0	50	10	4	TRUE
201	0	50	10	4	TRUE
200	0	50	10	4	TRUE
199	0	50	10	4	TRUE
198	0	50	10	4	TRUE
197	0	50	10	4	TRUE
196	0	50	10	4	TRUE
195	26	50	10	4	FALSE
194	17	42	10	4	FALSE
193	0	50	10	4	TRUE
192	0	50	10	12	TRUE
191	0	50	10	4	TRUE
190	0	50	10	4	TRUE
189	28	48	10	4	FALSE
188	0	50	10	4	TRUE
187	0	500	10	10	TRUE
186	0	500	10	10	TRUE
185	5	469	10	10	FALSE
184	8	465	10	10	FALSE
183	8	485	10	10	FALSE
182	5	464	10	10	FALSE
181	0	500	10	10	TRUE
180	0	500	10	10	TRUE
179	0	500	10	10	TRUE
178	0	500	10	10	TRUE
177	6	473	10	10	FALSE
176	0	500	10	10	TRUE
175	0	500	10	10	TRUE
174	0	500	10	10	TRUE
173	0	500	10	10	TRUE
172	0	500	10	10	TRUE
171	0	500	10	10	TRUE

Id	Number Of Found Patterns	Number Of Edit Scripts	Minimum Node Count	Minimum Support	Exception Occured
170	0	50	30	3	TRUE
169	3	50	30	3	FALSE
168	0	50	30	3	TRUE
167	0	50	30	3	TRUE
166	0	50	30	3	TRUE
165	0	44	30	3	FALSE
164	3	50	30	3	FALSE
163	0	50	30	3	TRUE
162	0	50	30	3	TRUE
161	0	50	30	3	TRUE
160	0	50	30	3	TRUE
159	2	45	30	3	FALSE
158	0	50	30	3	TRUE
157	0	50	30	3	TRUE
156	0	50	30	3	TRUE
155	0	50	30	3	TRUE
154	0	50	30	3	TRUE
153	0	50	30	3	TRUE
152	0	50	30	3	TRUE
151	0	50	30	3	TRUE
150	0	50	30	3	TRUE
149	1	45	30	3	FALSE
148	0	50	30	3	TRUE
147	3	49	30	3	FALSE
146	0	50	30	3	TRUE
145	4	44	30	3	FALSE
144	0	50	30	3	TRUE
143	0	50	30	3	TRUE
142	0	50	30	3	TRUE
141	0	50	30	3	TRUE
140	0	50	30	3	TRUE
139	1	47	30	3	FALSE
138	0	50	30	3	TRUE
137	0	50	30	3	TRUE
136	0	50	30	3	TRUE
135	0	50	30	3	TRUE
134	0	50	30	3	TRUE
133	2	47	30	3	FALSE
132	0	2000	10	8	TRUE

Id	Number Of Found Patterns	Number Of Edit Scripts	Minimum Node Count	Minimum Support	Exception Occured
131	13	180	10	8	FALSE
130	0	200	10	8	TRUE
129	12	189	10	8	FALSE
128	13	191	10	8	FALSE
127	0	200	10	8	TRUE
126	0	200	10	8	TRUE
125	10	188	10	8	FALSE
124	0	200	10	8	TRUE
123	0	200	10	8	TRUE
122	0	200	10	8	TRUE
121	0	200	10	8	TRUE
120	0	200	10	8	TRUE
119	0	200	10	8	TRUE
118	18	181	10	8	FALSE
117	0	200	10	8	TRUE
116	0	200	10	8	TRUE
115	12	182	10	8	FALSE
114	17	187	10	8	FALSE
113	0	200	10	8	TRUE
112	10	193	10	8	FALSE
111	13	193	10	8	FALSE
110	0	200	10	7	TRUE
109	0	200	10	6	TRUE
108	0	200	10	6	TRUE
107	0	200	10	6	TRUE
106	0	200	10	6	TRUE
105	0	200	10	6	TRUE
104	8	197	10	6	FALSE
103	0	200	10	6	TRUE
102	0	200	10	6	TRUE
101	10	171	10	7	FALSE
100	0	200	10	7	TRUE
99	0	200	10	7	TRUE
98	10	182	10	7	FALSE
97	8	191	10	8	FALSE
96	9	187	10	8	FALSE
95	13	96	10	5	FALSE
94	0	125	10	5	TRUE
93	0	250	10	5	TRUE

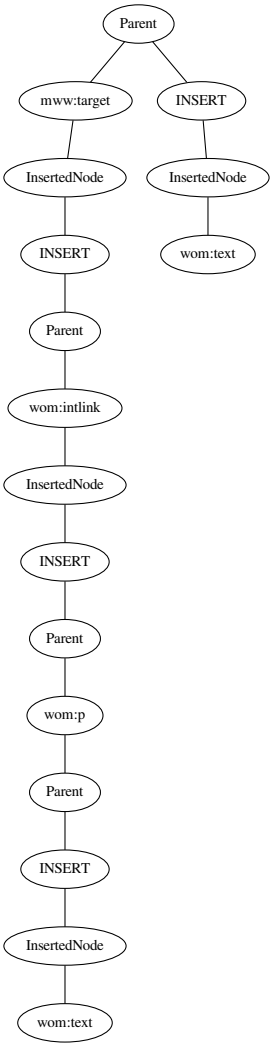
Id	Number Of Found Patterns	Number Of Edit Scripts	Minimum Node Count	Minimum Support	Exception Occured
92	0	250	10	5	TRUE
91	0	250	10	5	TRUE
90	0	500	10	5	TRUE
89	0	500	10	5	TRUE
88	0	500	10	5	TRUE
87	0	500	10	5	TRUE
86	0	500	10	5	TRUE
85	0	500	10	5	TRUE
84	0	500	10	5	TRUE
83	20	49	10	5	FALSE
82	0	50	10	5	TRUE
81	0	50	10	5	TRUE
80	14	45	10	5	FALSE
79	12	50	10	5	FALSE
78	4	50	10	10	FALSE
77	0	500	10	10	TRUE
76	4	474	10	11	FALSE
75	4	474	10	12	FALSE
74	0	474	20	12	FALSE
73	0	500	20	10	TRUE
72	0	500	20	8	TRUE
71	0	500	20	2	TRUE
70	0	500	30	1	TRUE
69	4	474	10	12	FALSE
68	0	1050	10	12	TRUE
67	0	550	10	12	TRUE
66	0	550	20	1	TRUE
65	0	1050	30	1	TRUE
64	0	1050	20	1	TRUE
63	0	1050	10	1	TRUE
62	0	50	10	1	TRUE
61	0	50	10	12	TRUE
60	0	50	10	12	TRUE
59	0	50	10	12	TRUE
58	1	50	10	12	FALSE
57	0	50	10	12	TRUE
56	0	50	10	12	TRUE
55	23	499	10	12	FALSE
54	0	500	10	12	TRUE

Id	Number Of Found Patterns	Number Of Edit Scripts	Minimum Node Count	Minimum Support	Exception Occured
53	0	500	10	12	TRUE
52	0	500	10	12	TRUE
51	0	500	10	12	TRUE
50	0	500	20	6	TRUE
49	0	500	20	6	TRUE
48	0	500	20	6	TRUE
47	0	1500	10	12	TRUE
46	0	1500	10	12	TRUE
45	0	1500	10	12	TRUE
44	17	499	10	12	FALSE
43	17	499	10	12	FALSE
42	0	500	10	12	TRUE
41	0	500	10	12	TRUE
40	0	500	10	8	TRUE
39	0	500	10	8	TRUE
38	0	500	10	8	TRUE
37	0	500	10	8	TRUE
36	0	1000	10	8	TRUE
35	0	1000	10	8	TRUE
34	0	10000	10	8	TRUE
33	0	10000	20	1	TRUE
32	0	10000	10	5	TRUE
31	0	10000	10	5	TRUE
30	0	10000	10	5	TRUE
29	0	10000	16	5	TRUE
28	0	10000	16	5	TRUE
27	0	10000	16	5	TRUE
26	0	10000	16	5	TRUE
25	0	10000	10	5	TRUE
24	0	10000	10	12	TRUE
23	0	10000	10	12	TRUE
22	0	10000	10	12	TRUE
21	0	50000	10	12	TRUE
20	0	50000	10	12	TRUE
19	0	50000	10	12	TRUE
18	0	100000	10	12	TRUE
17	0	100000	10	12	TRUE
16	0	100000	10	12	TRUE
15	0	100000	10	12	TRUE

Id	Number Of Found Patterns	Number Of Edit Scripts	Minimum Node Count	Minimum Support	Exception Occured
14	0	100000	10	12	TRUE
13	0	100000	10	12	TRUE
12	0	100000	10	12	TRUE
11	0	100000	10	12	TRUE
10	0	100000	10	12	TRUE
9	0	100000	10	12	TRUE
8	0	100000	10	12	TRUE
7	0	100000	10	12	TRUE
6	0	100000	10	12	TRUE
5	0	100000	10	12	TRUE
4	0	100000	10	12	TRUE
3	0	50	10	12	TRUE
2	0	50	10	12	TRUE
1	9	995	10	12	FALSE

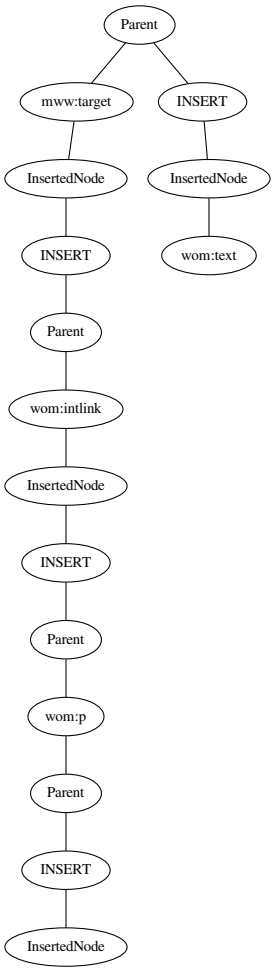
Appendix D Selected Frequent Pattern Graphs

Pattern No. 3



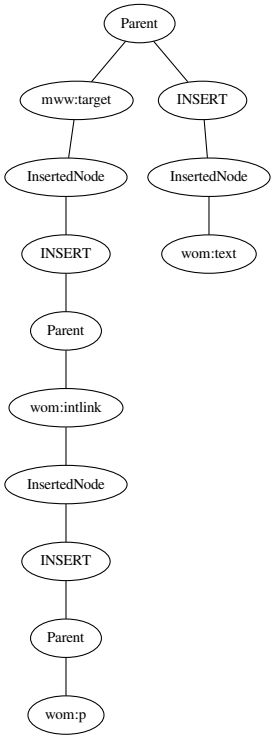
graph has occurrences in (at least) the following revision pairs :	revision A id: 12033 revision B id: 12046; revision A id: 98746 revision B id: 98749; revision A id: 175239 revision B id: 175260; revision A id: 279188 revision B id: 279189; revision A id: 282193 revision B id: 282194;
graph is subgraph of patterns :	10 34 43 44 50 51 52 74 75 80 81 87 88 89 90 114 116 119 120 121 122 142 143 144 145 146 147 148 149 150 166 173 174 175 176 177 178 220 221 222 223 224 237 238 239 240 241 267 271 272 278 279 297
graph is supergraph of patterns :	

Pattern No. 4


Parent
mww:target
INSERT
InsertedNode
InsertedNode
INSERT
wom:text
Parent
wom:intlink
InsertedNode
INSERT
Parent
wom:p
Parent
INSERT
InsertedNode

graph has occurrences in (at least) the following revision pairs :	revision A id: 12033 revision B id: 12046; revision A id: 98746 revision B id: 98749; revision A id: 175239 revision B id: 175260; revision A id: 279188 revision B id: 279189; revision A id: 282193 revision B id: 282194;
graph is subgraph of patterns :	10 11 34 43 44 50 51 52 53 54 55 56 74 75 79 80 81 85 87 88 89 90 91 114 116 119 120 121 122 142 143 144 145 146 147 148 149 150 166 167 173 174 175 176 177 178 220 221 222 223 224 225 237 238 239 240 241 267 271 272 278 279 280 281 282 297
graph is supergraph of patterns :	

Pattern No. 5



graph has occurrences in (at least) the following revision pairs :	revision A id: 12033 revision B id: 12046; revision A id: 98746 revision B id: 98749; revision A id: 175239 revision B id: 175260; revision A id: 279188 revision B id: 279189; revision A id: 282193 revision B id: 282194;
graph is subgraph of patterns :	10 11 34 43 44 50 51 52 53 54 55 56 74 75 79 80 81 85 87 88 89 90 91 92 111 112 113 114 115 116 119 120 121 122 137 142 143 144 145 146 147 148 149 150 166 167 173 174 175 176 177 178 203 220 221 222 223 224 225 237 238 239 240 241 244 245 267 271 272 278 279 280 281 282 297
graph is supergraph of patterns :	

Pattern No. 6

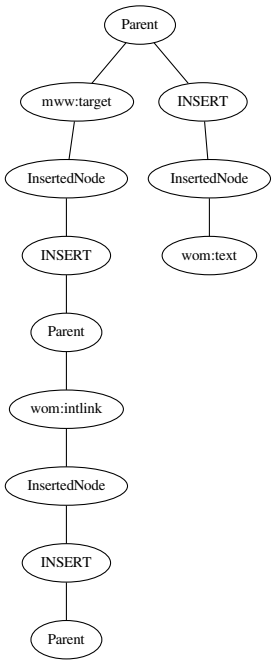
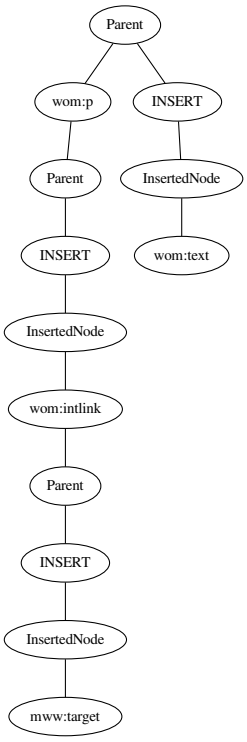


Figure 2.15: Pattern no. 6

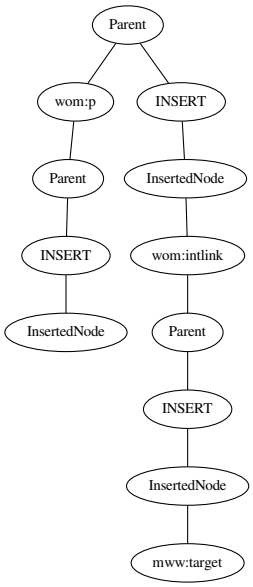
graph has occurrences in (at least) the following revision pairs :	revision A id: 12033 revision B id: 12046; revision A id: 58452 revision B id: 217584; revision A id: 98746 revision B id: 98749; revision A id: 108099 revision B id: 405282; revision A id: 108224 revision B id: 108236;
graph is subgraph of patterns :	10 11 12 20 21 24 25 34 35 41 42 43 44 50 51 52 53 54 55 56 59 60 74 75 76 79 80 81 82 83 85 87 88 89 90 91 92 93 111 112 113 114 115 116 117 119 120 121 122 123 124 135 137 142 143 144 145 146 147 148 149 150 151 164 166 167 168 173 174 175 176 177 178 184 185 192 203 209 220 221 222 223 224 225 226 227 228 235 236 237 238 239 240 241 244 245 246 247 248 267 271 272 278 279 280 281 282 285 289 297 299 300 301 307
graph is supergraph of patterns :	

Pattern No. 7



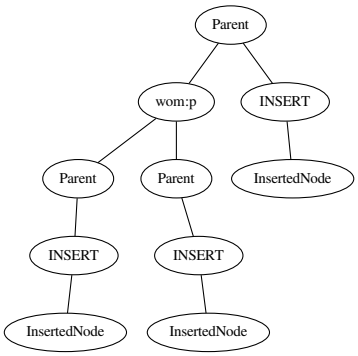
graph has occurrences in (at least) the following revision pairs :	revision A id: 12033 revision B id: 12046; revision A id: 98746 revision B id: 98749; revision A id: 175239 revision B id: 175260; revision A id: 279188 revision B id: 279189; revision A id: 282193 revision B id: 282194;
graph is subgraph of patterns :	10 14 34 43 44 50 51 52 69 74 75 77 80 81 87 88 89 90 114 116 119 120 121 122 142 143 144 145 146 147 148 149 150 166 173 174 175 176 177 178 220 221 222 223 224 237 238 239 240 241 253 254 267 271 272 278 279 297
graph is supergraph of patterns :	

Pattern No. 8



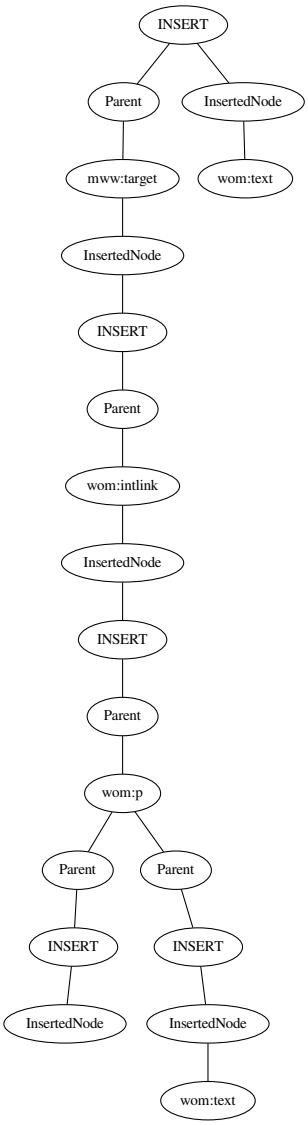
graph has occurrences in (at least) the following revision pairs :	revision A id: 12033 revision B id: 12046; revision A id: 98746 revision B id: 98749; revision A id: 175239 revision B id: 175260; revision A id: 279188 revision B id: 279189; revision A id: 282193 revision B id: 282194;
graph is subgraph of patterns :	10 11 14 18 34 43 44 50 51 52 53 54 55 56 69 74 75 77 79 80 81 85 87 88 89 90 91 114 116 119 120 121 122 142 143 144 145 146 147 148 149 150 166 167 173 174 175 176 177 178 220 221 222 223 224 225 237 238 239 240 241 253 254 259 260 267 271 272 278 279 280 281 282 295 297
graph is supergraph of patterns :	

Pattern No. 9



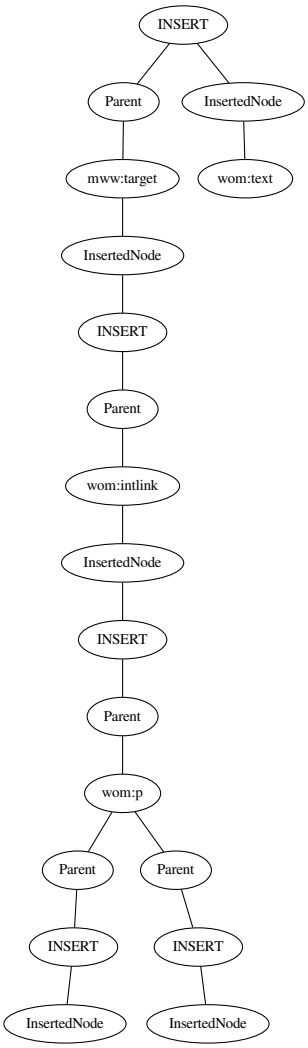
graph has occurrences in (at least) the following revision pairs :	revision A id: 12033 revision B id: 12046; revision A id: 175239 revision B id: 175260; revision A id: 308634 revision B id: 308661; revision A id: 332508 revision B id: 686549; revision A id: 359376 revision B id: 359392;
graph is subgraph of patterns :	10 11 13 14 15 18 43 44 51 52 57 65 68 69 75 77 79 80 81 85 87 88 89 90 116 119 120 121 122 125 126 142 143 144 145 146 147 148 149 150 153 154 155 166 167 169 170 173 174 175 176 177 178 194 195 202 205 207 220 221 222 225 237 238 239 240 241 249 250 251 252 253 254 267 271 272 274 275 276 278 279 297
graph is supergraph of patterns :	

Pattern No. 10



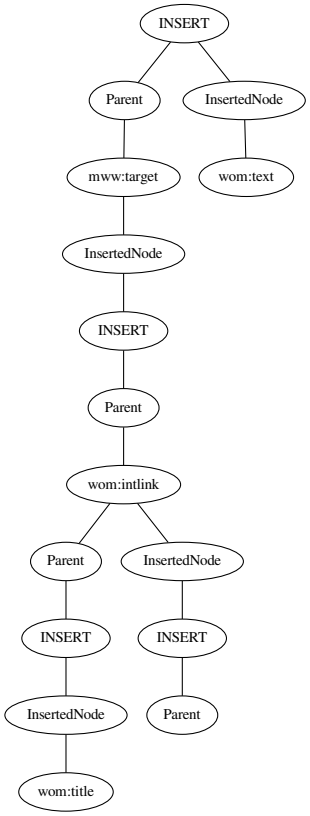
graph has occurrences in (at least) the following revision pairs :	revision A id: 12033 revision B id: 12046; revision A id: 175239 revision B id: 175260; revision A id: 308634 revision B id: 308661; revision A id: 332508 revision B id: 686549; revision A id: 359376 revision B id: 359392;
graph is subgraph of patterns :	43 44 51 52 75 80 81 87 88 89 90 116 119 120 121 122 142 143 144 145 146 147 148 149 150 166 173 174 175 176 177 178 220 221 222 237 238 239 240 241 267 271 272 278 279 297
graph is supergraph of patterns :	3 4 5 6 7 8 9 11 14 15 18

Pattern No. 11



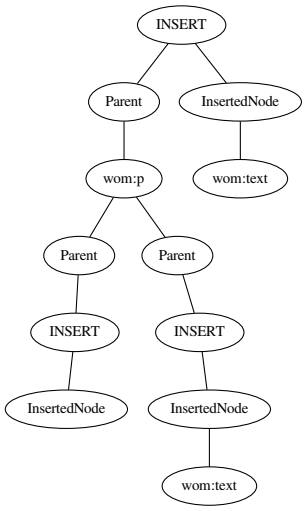
graph has occurrences in (at least) the following revision pairs :	revision A id: 12033 revision B id: 12046; revision A id: 175239 revision B id: 175260; revision A id: 308634 revision B id: 308661; revision A id: 332508 revision B id: 686549; revision A id: 359376 revision B id: 359392;
graph is subgraph of patterns :	10 43 44 51 52 75 79 80 81 85 87 88 89 90 116 119 120 121 122 142 143 144 145 146 147 148 149 150 166 167 173 174 175 176 177 178 220 221 222 225 237 238 239 240 241 267 271 272 278 279 297
graph is supergraph of patterns :	4 5 6 8 9

Pattern No. 12



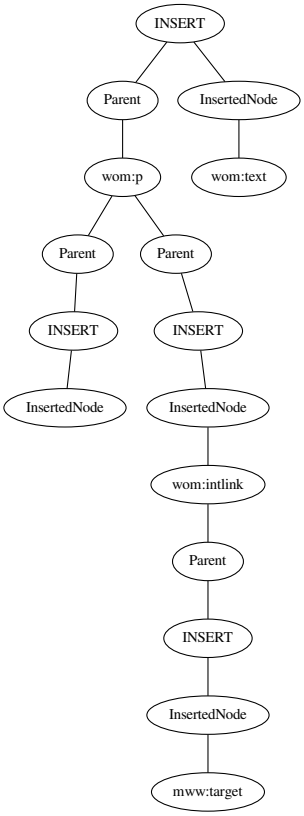
graph has occurrences in (at least) the following revision pairs :	revision A id: 12033 revision B id: 12046; revision A id: 108224 revision B id: 108236; revision A id: 175239 revision B id: 175260; revision A id: 279188 revision B id: 279189; revision A id: 308634 revision B id: 308661;
graph is subgraph of patterns :	20 21 41 50 74 79 87 89 90 91 92 112 117 121 122 123 124 137 142 143 144 145 146 147 164 174 177 178 184 192 203 220 225 227 228 246 248 272 278 279 280 281 282 285 297 299 300 301 307
graph is supergraph of patterns :	6

Pattern No. 13



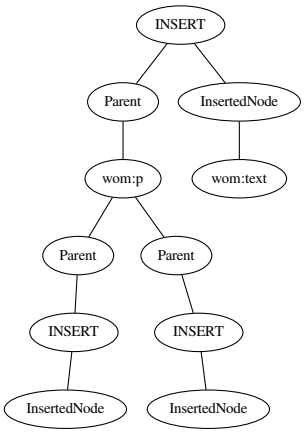
graph has occurrences in (at least) the following revision pairs :	revision A id: 12033 revision B id: 12046; revision A id: 175239 revision B id: 175260; revision A id: 308634 revision B id: 308661; revision A id: 332508 revision B id: 686549; revision A id: 359376 revision B id: 359392;
graph is subgraph of patterns :	43 44 51 52 57 69 75 80 81 87 88 89 90 116 119 120 121 122 125 126 142 143 144 145 146 147 148 149 150 153 154 155 166 169 170 173 174 175 176 177 178 194 195 202 237 238 239 240 241 249 250 251 252 267 271 272 274 275 276 297
graph is supergraph of patterns :	9

Pattern No. 14



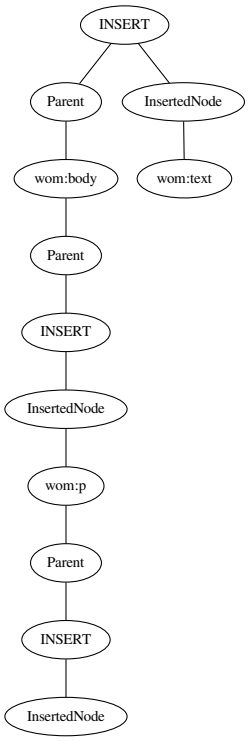
graph has occurrences in (at least) the following revision pairs :	revision A id: 12033 revision B id: 12046; revision A id: 175239 revision B id: 175260; revision A id: 308634 revision B id: 308661; revision A id: 332508 revision B id: 686549; revision A id: 359376 revision B id: 359392;
graph is subgraph of patterns :	10 43 44 51 52 69 75 77 80 81 87 88 89 90 116 119 120 121 122 142 143 144 145 146 147 148 149 150 166 173 174 175 176 177 178 220 221 222 237 238 239 240 241 253 254 267 271 272 278 279 297
graph is supergraph of patterns :	7 8 9

Pattern No. 15



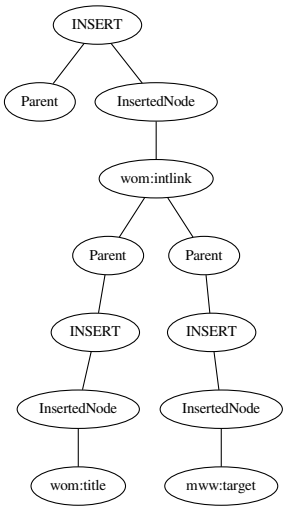
graph has occurrences in (at least) the following revision pairs :	revision A id: 12033 revision B id: 12046; revision A id: 175239 revision B id: 175260; revision A id: 308634 revision B id: 308661; revision A id: 332508 revision B id: 686549; revision A id: 359376 revision B id: 359392;
graph is subgraph of patterns :	10 43 44 51 52 57 65 69 75 77 80 81 87 88 89 90 116 119 120 121 122 125 126 142 143 144 145 146 147 148 149 150 153 154 155 166 169 170 173 174 175 176 177 178 194 195 202 220 221 222 237 238 239 240 241 249 250 251 252 253 254 267 271 272 274 275 276 278 279 297
graph is supergraph of patterns :	9

Pattern No. 16



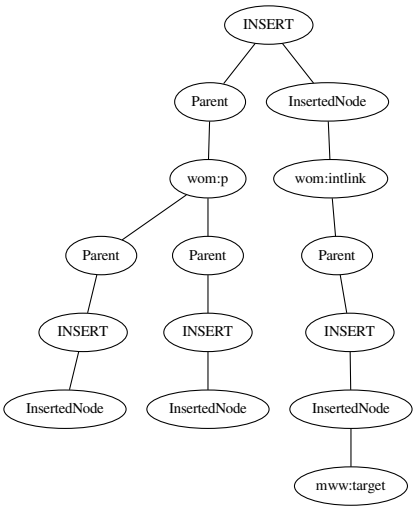
graph has occurrences in (at least) the following revision pairs :	revision A id: 12033 revision B id: 12046; revision A id: 98746 revision B id: 98749; revision A id: 108099 revision B id: 405282; revision A id: 175239 revision B id: 175260; revision A id: 279562 revision B id: 279563;
graph is subgraph of patterns :	26 34 36 61 70 72 73 74 89 111 112 113 119 120 125 143 145 146 154 173 174 175 176 180 181 193 200 201 234 242 244 249 250 251 290 309 310
graph is supergraph of patterns :	

Pattern No. 17



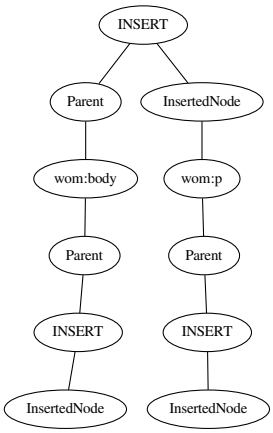
graph has occurrences in (at least) the following revision pairs :	revision A id: 12033 revision B id: 12046; revision A id: 108224 revision B id: 108236; revision A id: 175239 revision B id: 175260; revision A id: 279188 revision B id: 279189; revision A id: 308634 revision B id: 308661;
graph is subgraph of patterns :	20 21 22 23 41 50 74 79 87 89 90 91 92 112 117 121 122 123 124 137 142 143 144 145 146 147 164 174 177 178 184 192 203 220 225 227 228 246 248 272 278 279 280 281 282 283 284 285 297 299 300 301 307
graph is supergraph of patterns :	

Pattern No. 18



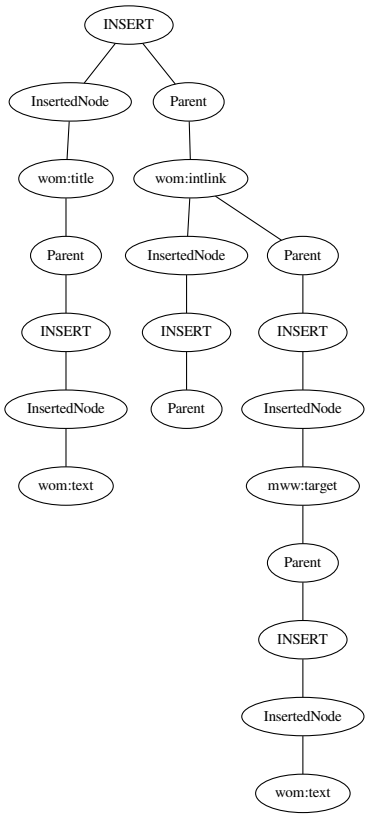
graph has occurrences in (at least) the following revision pairs :	revision A id: 12033 revision B id: 12046; revision A id: 175239 revision B id: 175260; revision A id: 308634 revision B id: 308661; revision A id: 332508 revision B id: 686549; revision A id: 359376 revision B id: 359392;
graph is subgraph of patterns :	10 43 44 51 52 69 75 77 79 80 81 85 87 88 89 90 116 119 120 121 122 142 143 144 145 146 147 148 149 150 166 167 173 174 175 176 177 178 220 221 222 225 237 238 239 240 241 253 254 267 271 272 278 279 297
graph is supergraph of patterns :	8 9

Pattern No. 19



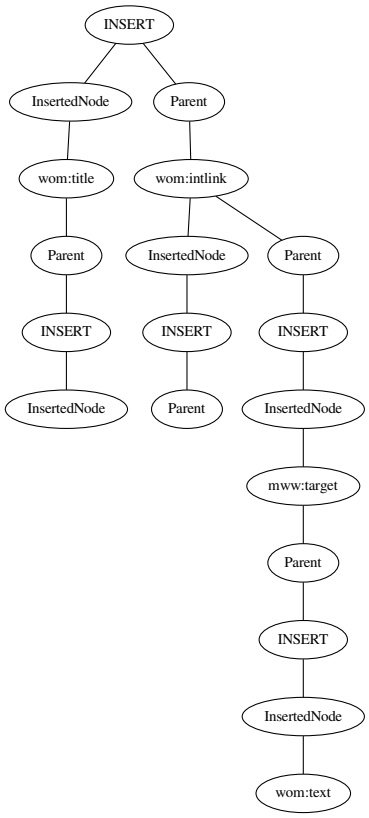
graph has occurrences in (at least) the following revision pairs :	revision A id: 12033 revision B id: 12046; revision A id: 98746 revision B id: 98749; revision A id: 108099 revision B id: 405282; revision A id: 175239 revision B id: 175260; revision A id: 279562 revision B id: 279563;
graph is subgraph of patterns :	26 34 36 61 66 70 72 73 74 89 111 112 113 119 120 125 143 145 146 154 173 174 175 176 180 181 193 200 201 234 242 244 249 250 251 290 309 310
graph is supergraph of patterns :	

Pattern No. 20



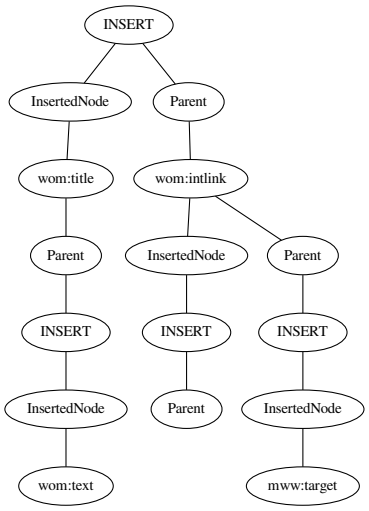
graph has occurrences in (at least) the following revision pairs :	revision A id: 12033 revision B id: 12046; revision A id: 108224 revision B id: 108236; revision A id: 175239 revision B id: 175260; revision A id: 279188 revision B id: 279189; revision A id: 308634 revision B id: 308661;
graph is subgraph of patterns :	41 50 74 79 87 89 90 91 112 117 121 122 123 124 137 142 164 177 184 192 225 228 248 272 278 279 285 297 299 300 301 307
graph is supergraph of patterns :	6 12 17 21 22 23

Pattern No. 21



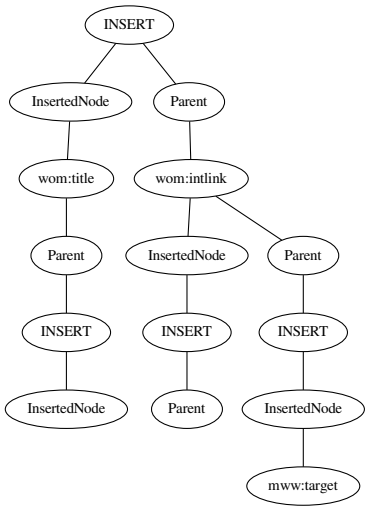
graph has occurrences in (at least) the following revision pairs :	revision A id: 12033 revision B id: 12046; revision A id: 108224 revision B id: 108236; revision A id: 175239 revision B id: 175260; revision A id: 279188 revision B id: 279189; revision A id: 308634 revision B id: 308661;
graph is subgraph of patterns :	20 41 50 74 79 87 89 90 91 112 117 121 122 123 124 137 142 143 144 145 146 147 164 174 177 178 184 192 225 228 246 248 272 278 279 285 297 299 300 301 307
graph is supergraph of patterns :	6 12 17

Pattern No. 22



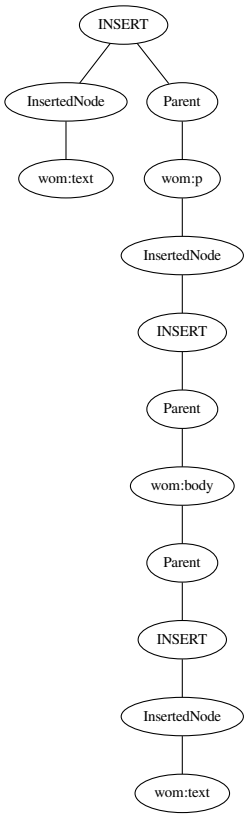
graph has occurrences in (at least) the following revision pairs :	revision A id: 12033 revision B id: 12046; revision A id: 108224 revision B id: 108236; revision A id: 175239 revision B id: 175260; revision A id: 279188 revision B id: 279189; revision A id: 308634 revision B id: 308661;
graph is subgraph of patterns :	20 41 50 74 79 87 89 90 91 112 117 121 122 123 124 137 142 164 177 184 192 225 228 248 272 278 279 285 297 299 300 301 307
graph is supergraph of patterns :	17

Pattern No. 23



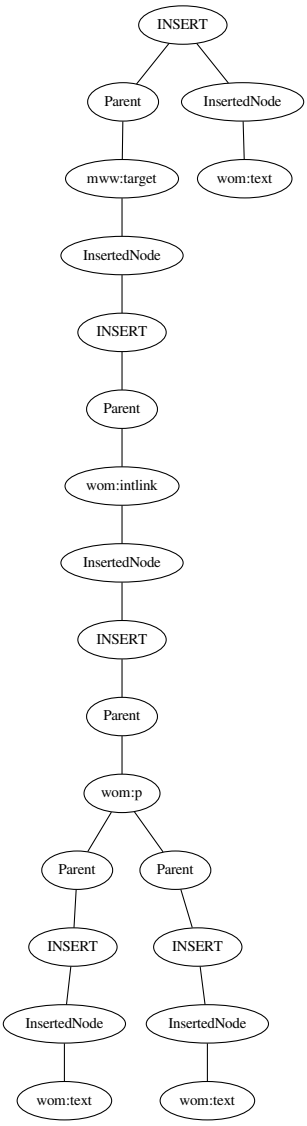
graph has occurrences in (at least) the following revision pairs :	revision A id: 12033 revision B id: 12046; revision A id: 108224 revision B id: 108236; revision A id: 175239 revision B id: 175260; revision A id: 279188 revision B id: 279189; revision A id: 308634 revision B id: 308661;
graph is subgraph of patterns :	20 41 50 74 79 87 89 90 91 112 117 121 122 123 124 137 142 143 144 145 146 147 164 174 177 178 184 192 225 228 246 248 272 278 279 285 297 299 300 301 307
graph is supergraph of patterns :	17

Pattern No. 26



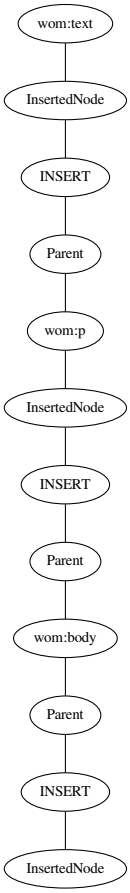
graph has occurrences in (at least) the following revision pairs :	revision A id: 12033 revision B id: 12046; revision A id: 98746 revision B id: 98749; revision A id: 108099 revision B id: 405282; revision A id: 175239 revision B id: 175260; revision A id: 279562 revision B id: 279563;
graph is subgraph of patterns :	34 36 61 70 72 73 74 89 119 120 125 143 145 146 154 173 174 175 176 180 181 193 200 201 242 249 250 251 290 309 310
graph is supergraph of patterns :	16 19 66

Pattern No. 44



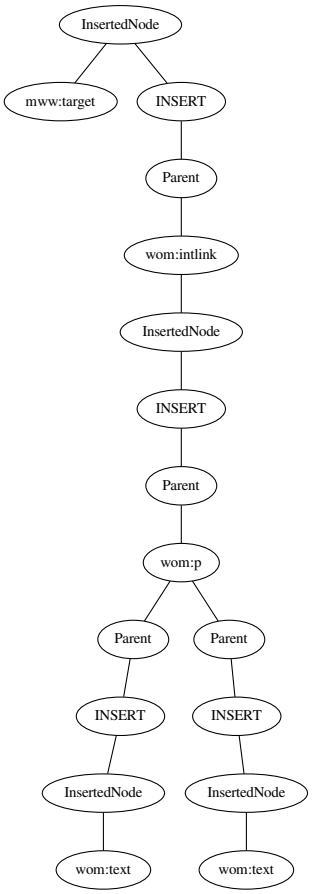
graph has occurrences in (at least) the following revision pairs :	revision A id: 12033 revision B id: 12046; revision A id: 175239 revision B id: 175260; revision A id: 308634 revision B id: 308661; revision A id: 332508 revision B id: 686549; revision A id: 359376 revision B id: 359392;
graph is subgraph of patterns :	51 52 75 80 81 87 88 89 90 116 119 120 121 122 142 143 144 145 146 147 148 149 150 166 173 174 175 176 177 178 237 238 239 240 241 267 271 272 297
graph is supergraph of patterns :	3 4 5 6 7 8 9 10 11 13 14 15 18 69

Pattern No. 66



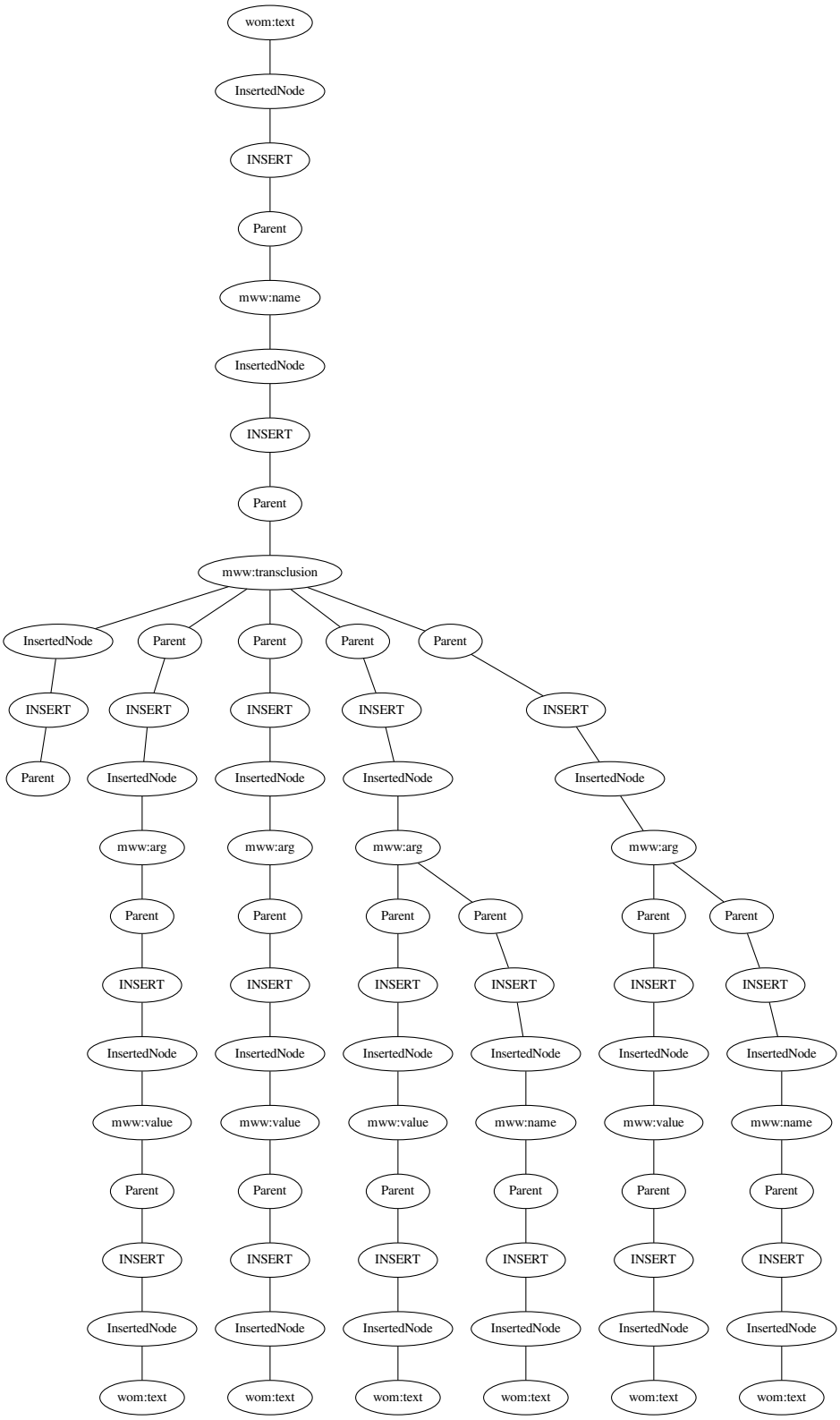
graph has occurrences in (at least) the following revision pairs :	revision A id: 12033 revision B id: 12046; revision A id: 98746 revision B id: 98749; revision A id: 108099 revision B id: 405282; revision A id: 175239 revision B id: 175260; revision A id: 279562 revision B id: 279563;
graph is subgraph of patterns :	26 34 36 61 70 72 73 74 89 119 120 125 143 145 146 154 173 174 175 176 180 181 193 200 201 242 249 250 251 290 309 310
graph is supergraph of patterns :	19

Pattern No. 69



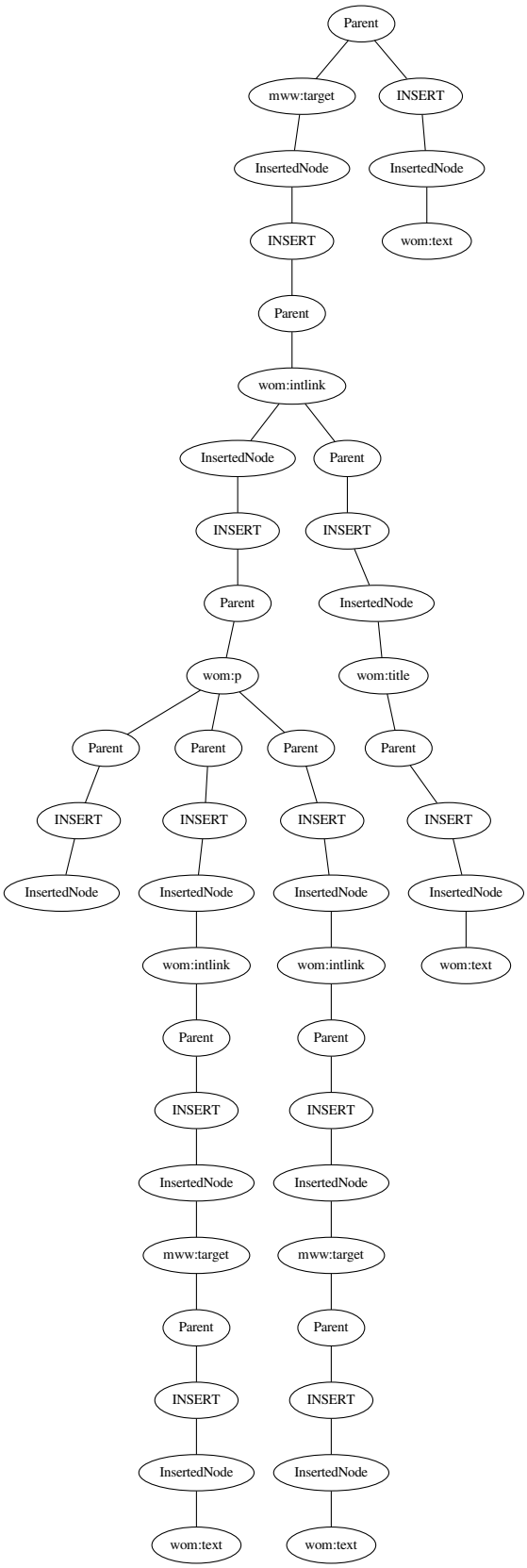
graph has occurrences in (at least) the following revision pairs :	revision A id: 12033 revision B id: 12046; revision A id: 175239 revision B id: 175260; revision A id: 308634 revision B id: 308661; revision A id: 332508 revision B id: 686549; revision A id: 359376 revision B id: 359392;
graph is subgraph of patterns :	43 44 51 52 75 80 81 87 88 89 90 116 119 120 121 122 142 143 144 145 146 147 148 149 150 166 173 174 175 176 177 178 237 238 239 240 241 267 271 272 297
graph is supergraph of patterns :	7 8 9 13 14 15 18

Pattern No. 78


wom:text
InsertedNode
INSERT
Parent
mww:name
InsertedNode
INSERT
Parent
mww:transclusion
InsertedNode
Parent
Parent
Parent
Parent
INSERT
INSERT
INSERT
INSERT
INSERT
Parent
InsertedNode
InsertedNode
InsertedNode
InsertedNode
mww:arg
mww:arg
mww:arg
mww:arg
Parent
Parent
Parent
Parent
Parent
Parent
INSERT
INSERT
INSERT
INSERT
INSERT
INSERT
InsertedNode
InsertedNode
InsertedNode
InsertedNode
InsertedNode
InsertedNode
mww:value
mww:value
mww:value
mww:name
mww:value
mww:name
Parent
Parent
Parent
Parent
Parent
Parent
INSERT
INSERT
INSERT
INSERT
INSERT
INSERT
InsertedNode
InsertedNode
InsertedNode
InsertedNode
InsertedNode
InsertedNode
wom:text
wom:text
wom:text
wom:text
wom:text
wom:text

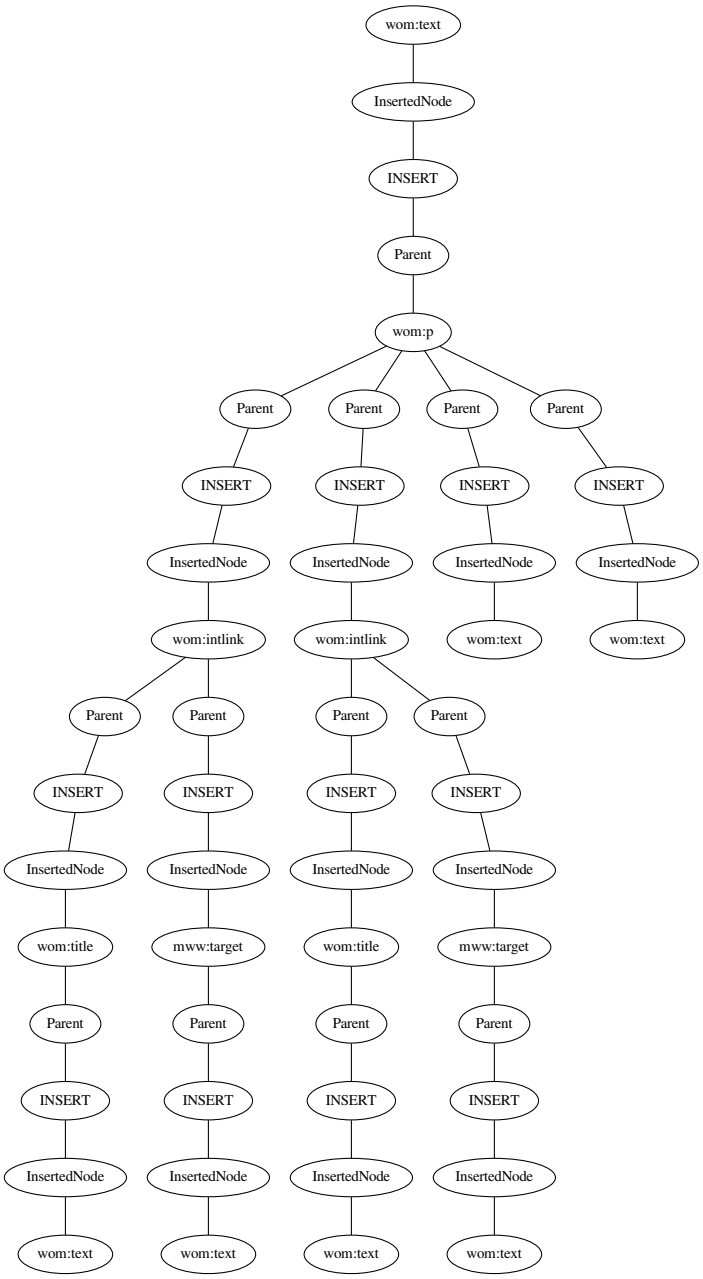
graph has occurrences in (at least) the following revision pairs :	revision A id: 90853654 revision B id: 106357619; revision A id: 149216395 revision B id: 149791818; revision A id: 181855521 revision B id: 188528728; revision A id: 268363914 revision B id: 268364383; revision A id: 281074108 revision B id: 281074807;
graph is subgraph of patterns :	
graph is supergraph of patterns :	1 2 38 84 95 172 188 189 302 314

Pattern No. 79


Parent
mww:target
INSERT
InsertedNode
InsertedNode
INSERT
wom:text
Parent
wom:intlink
InsertedNode
Parent
INSERT
INSERT
Parent
InsertedNode
wom:p
wom:title
Parent
Parent
Parent
Parent
INSERT
INSERT
INSERT
INSERT
InsertedNode
InsertedNode
InsertedNode
InsertedNode
wom:intlink
wom:intlink
wom:text
Parent
Parent
INSERT
INSERT
InsertedNode
InsertedNode
mww:target
mww:target
Parent
Parent
INSERT
INSERT
InsertedNode
InsertedNode
wom:text
wom:text

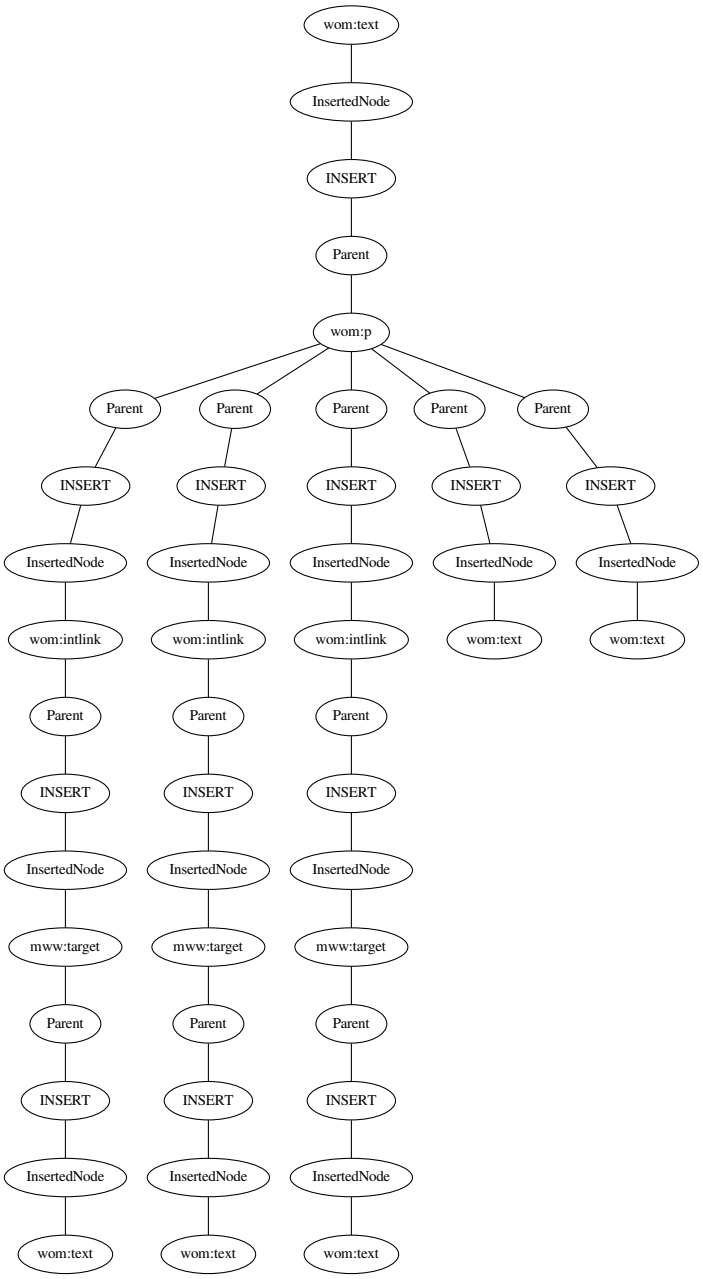
graph has occurrences in (at least) the following revision pairs :	revision A id: 175239 revision B id: 175260; revision A id: 308634 revision B id: 308661; revision A id: 359376 revision B id: 359392; revision A id: 621714 revision B id: 622105; revision A id: 1207572 revision B id: 1238684;
graph is subgraph of patterns :	142 225 278
graph is supergraph of patterns :	4 5 6 8 9 11 12 17 18 20 21 22 23 54 68 91 92 137 167 281 282 284

Pattern No. 87



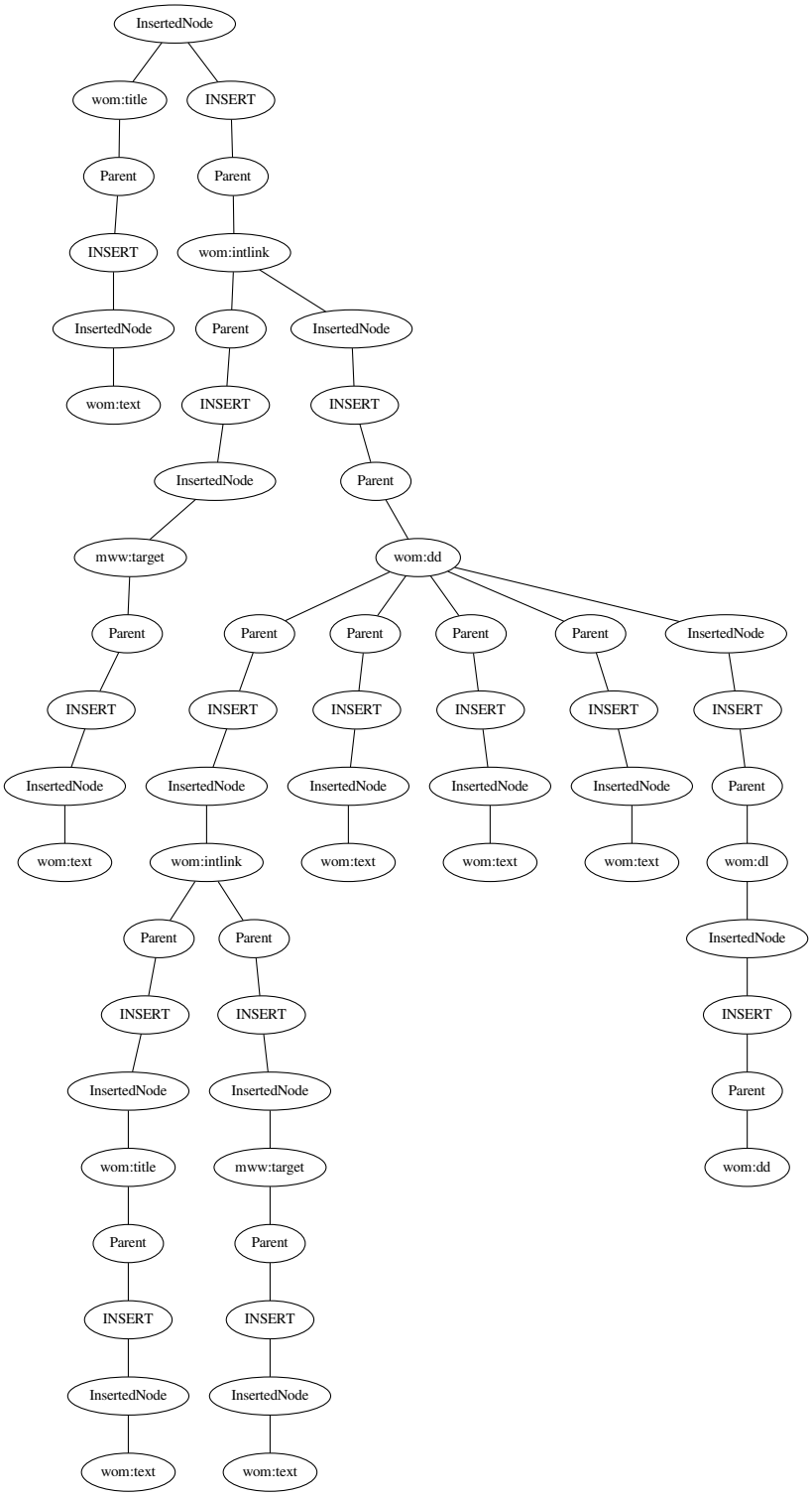
graph has occurrences in (at least) the following revision pairs :	revision A id: 3034654 revision B id: 3034706; revision A id: 5359016 revision B id: 8141976; revision A id: 61048870 revision B id: 61050239; revision A id: 73469485 revision B id: 73509280; revision A id: 77437862 revision B id: 77438215;
graph is subgraph of patterns :	121 142
graph is supergraph of patterns :	3 4 5 6 7 8 9 10 11 12 13 14 15 17 18 20 21 22 23 43 44 50 51 54 57 65 68 69 90 91 92 116 137 153 166 167 170 177 178 222 237 240 254 272 279 281 282 284

Pattern No. 88



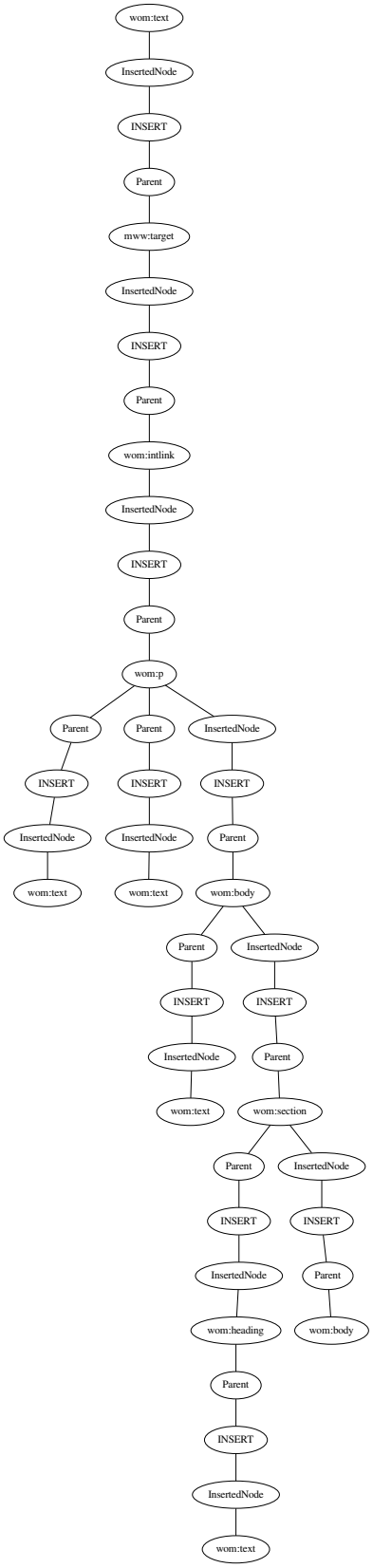
graph has occurrences in (at least) the following revision pairs :	revision A id: 1378554 revision B id: 1378561; revision A id: 4734004 revision B id: 4734726; revision A id: 5359016 revision B id: 8141976; revision A id: 7421939 revision B id: 7720063; revision A id: 8845074 revision B id: 9117591;
graph is subgraph of patterns :	142 148 149
graph is supergraph of patterns :	3 4 5 6 7 8 9 10 11 13 14 15 18 43 44 51 54 57 65 68 69 81 116 153 166 167 170 222 237 240 253 254 267

Pattern No. 117



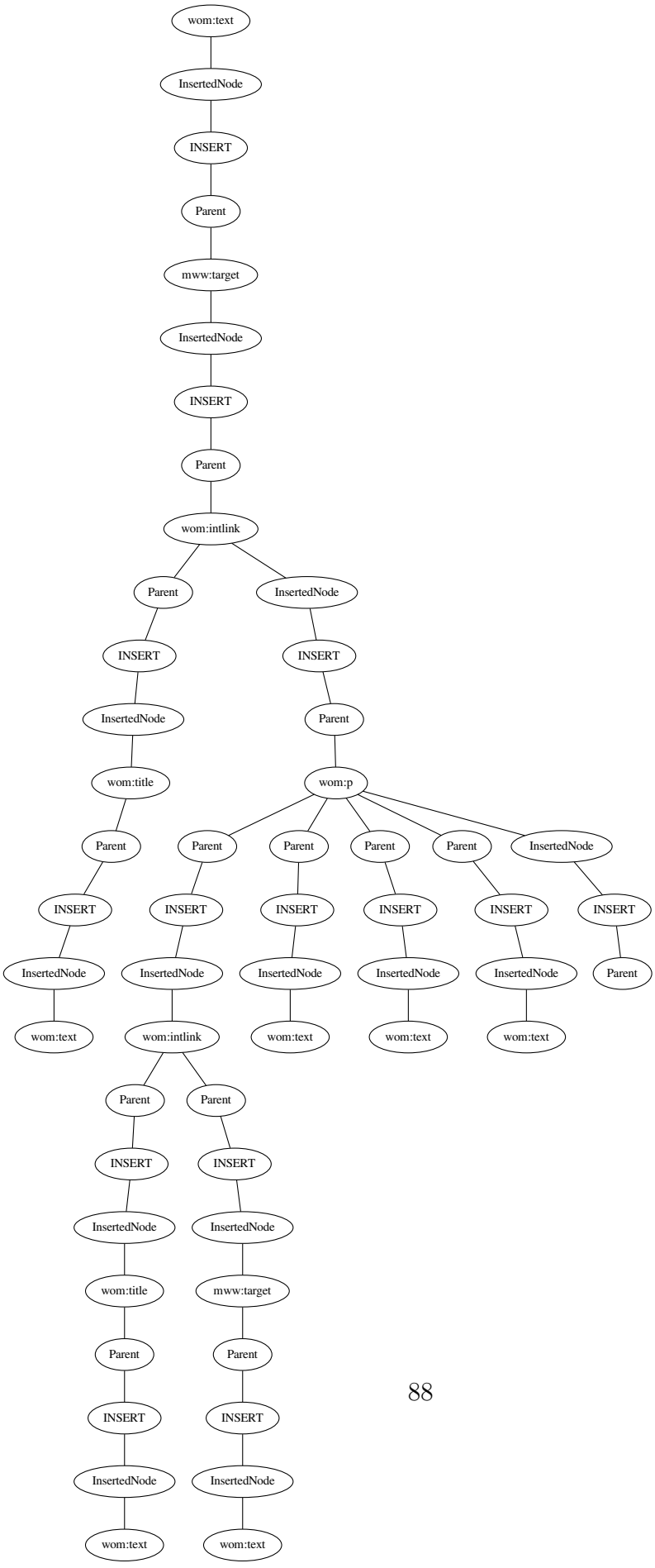
graph has occurrences in (at least) the following revision pairs :	revision A id: 3489847 revision B id: 4061708; revision A id: 73469485 revision B id: 73509280; revision A id: 103960624 revision B id: 104012474; revision A id: 125522905 revision B id: 125522953; revision A id: 262260060 revision B id: 262308574;
graph is subgraph of patterns :	
graph is supergraph of patterns :	6 12 17 20 21 22 23 24 25 27 28 29 30 31 32 33 124 192 209 215 217 258 285 300 301

Pattern No. 119


wom:text
InsertedNode
INSERT
Parent
mww:target
InsertedNode
INSERT
Parent
wom:intlink
InsertedNode
INSERT
Parent
wom:p
Parent
Parent
InsertedNode
INSERT
INSERT
INSERT
InsertedNode
InsertedNode
Parent
wom:text
wom:text
wom:body
Parent
InsertedNode
INSERT
INSERT
InsertedNode
Parent
wom:text
wom:section
Parent
InsertedNode
INSERT
INSERT
InsertedNode
Parent
wom:heading
wom:body
Parent
INSERT
InsertedNode
wom:text

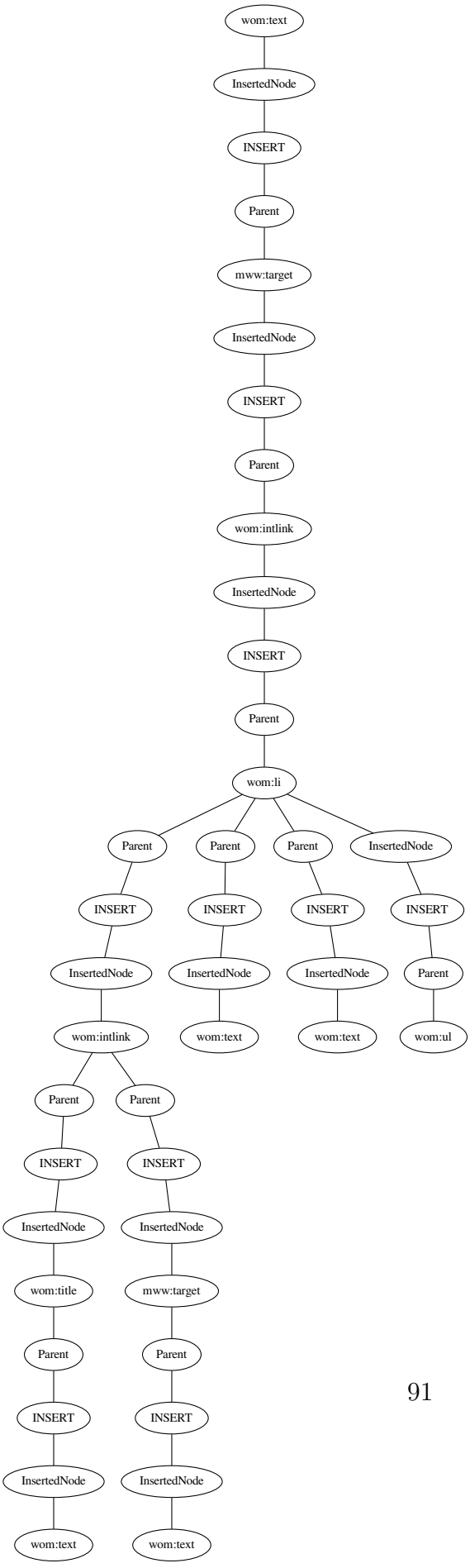
graph has occurrences in (at least) the following revision pairs :	revision A id: 3206207 revision B id: 3206266; revision A id: 4991103 revision B id: 5000801; revision A id: 5295032 revision B id: 5295067; revision A id: 5359016 revision B id: 8141976; revision A id: 5561215 revision B id: 5580959;
graph is subgraph of patterns :	145
graph is supergraph of patterns :	3 4 5 6 7 8 9 10 11 13 14 15 16 18 19 26 34 44 52 55 56 61 66 69 101 102 113 114 115 120 134 154 155 201 207 208 271 275 286 290

Pattern No. 121


wom:text
InsertedNode
INSERT
Parent
mww:target
InsertedNode
INSERT
Parent
wom:intlink
Parent
InsertedNode
INSERT
INSERT
InsertedNode
Parent
wom:title
wom:p
Parent
Parent
Parent
Parent
Parent
InsertedNode
INSERT
INSERT
INSERT
INSERT
INSERT
INSERT
InsertedNode
InsertedNode
InsertedNode
InsertedNode
InsertedNode
Parent
wom:text
wom:intlink
wom:text
wom:text
wom:text
Parent
Parent
INSERT
INSERT
InsertedNode
InsertedNode
wom:title
mww:target
Parent
Parent
INSERT
INSERT
InsertedNode
InsertedNode
wom:text
wom:text

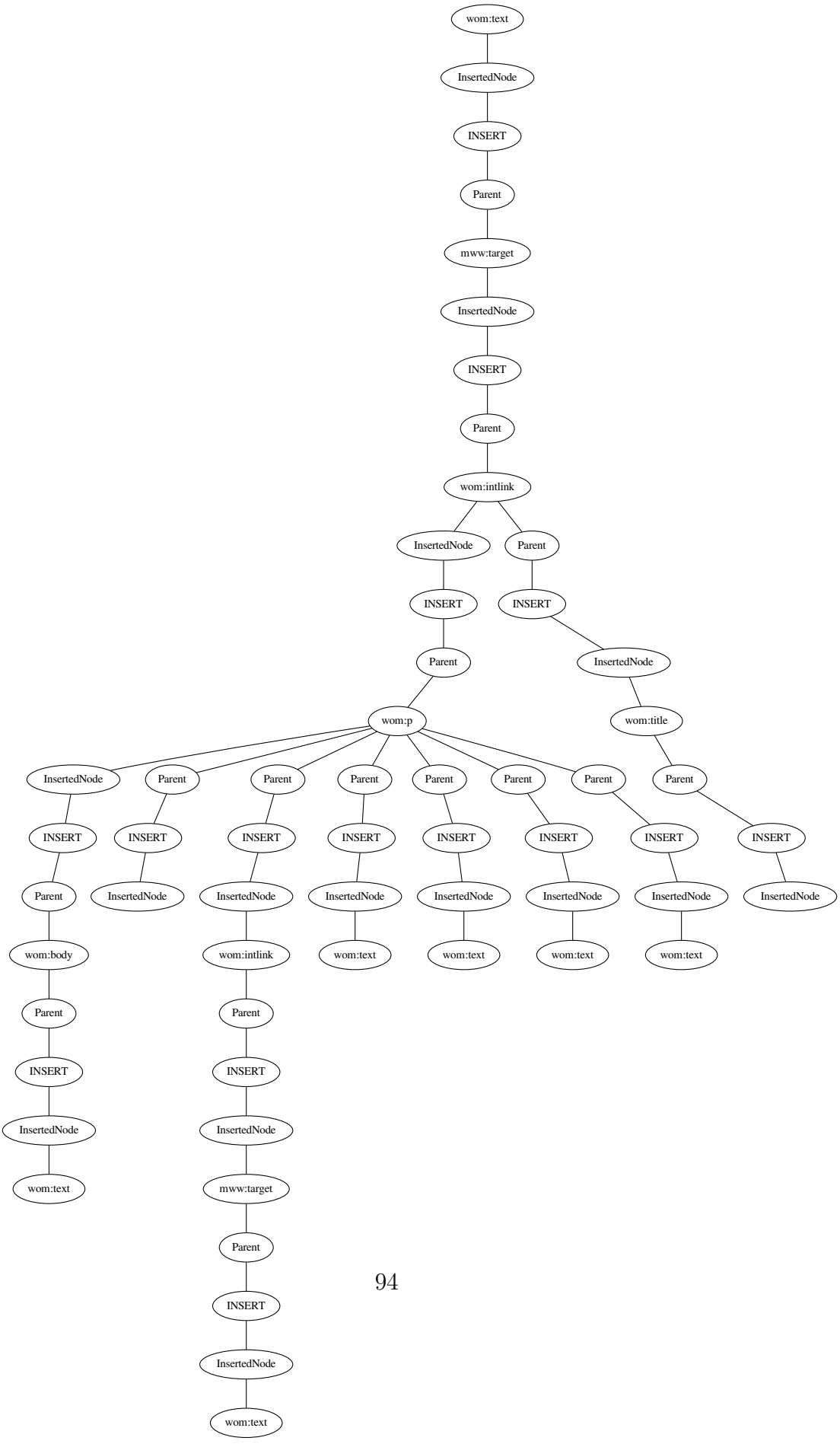
graph has occurrences in (at least) the following revision pairs :	revision A id: 3034654 revision B id: 3034706; revision A id: 5359016 revision B id: 8141976; revision A id: 73469485 revision B id: 73509280; revision A id: 116455033 revision B id: 116455138; revision A id: 150823988 revision B id: 172390487;
graph is subgraph of patterns :	142
graph is supergraph of patterns :	3 4 5 6 7 8 9 10 11 12 13 14 15 17 18 20 21 22 23 43 44 50 51 52 53 54 56 57 65 68 69 87 90 91 92 114 115 116 137 147 153 155 166 167 170 177 178 222 237 240 241 254 272 279 281 282 284 297

Pattern No. 123


wom:text
InsertedNode
INSERT
Parent
mww:target
InsertedNode
INSERT
Parent
wom:intlink
InsertedNode
INSERT
Parent
wom:li
Parent
Parent
Parent
InsertedNode
INSERT
INSERT
INSERT
INSERT
InsertedNode
InsertedNode
InsertedNode
Parent
wom:intlink
wom:text
wom:text
wom:ul
Parent
Parent
INSERT
INSERT
InsertedNode
InsertedNode
wom:title
mww:target
Parent
Parent
INSERT
INSERT
InsertedNode
InsertedNode
wom:text
wom:text

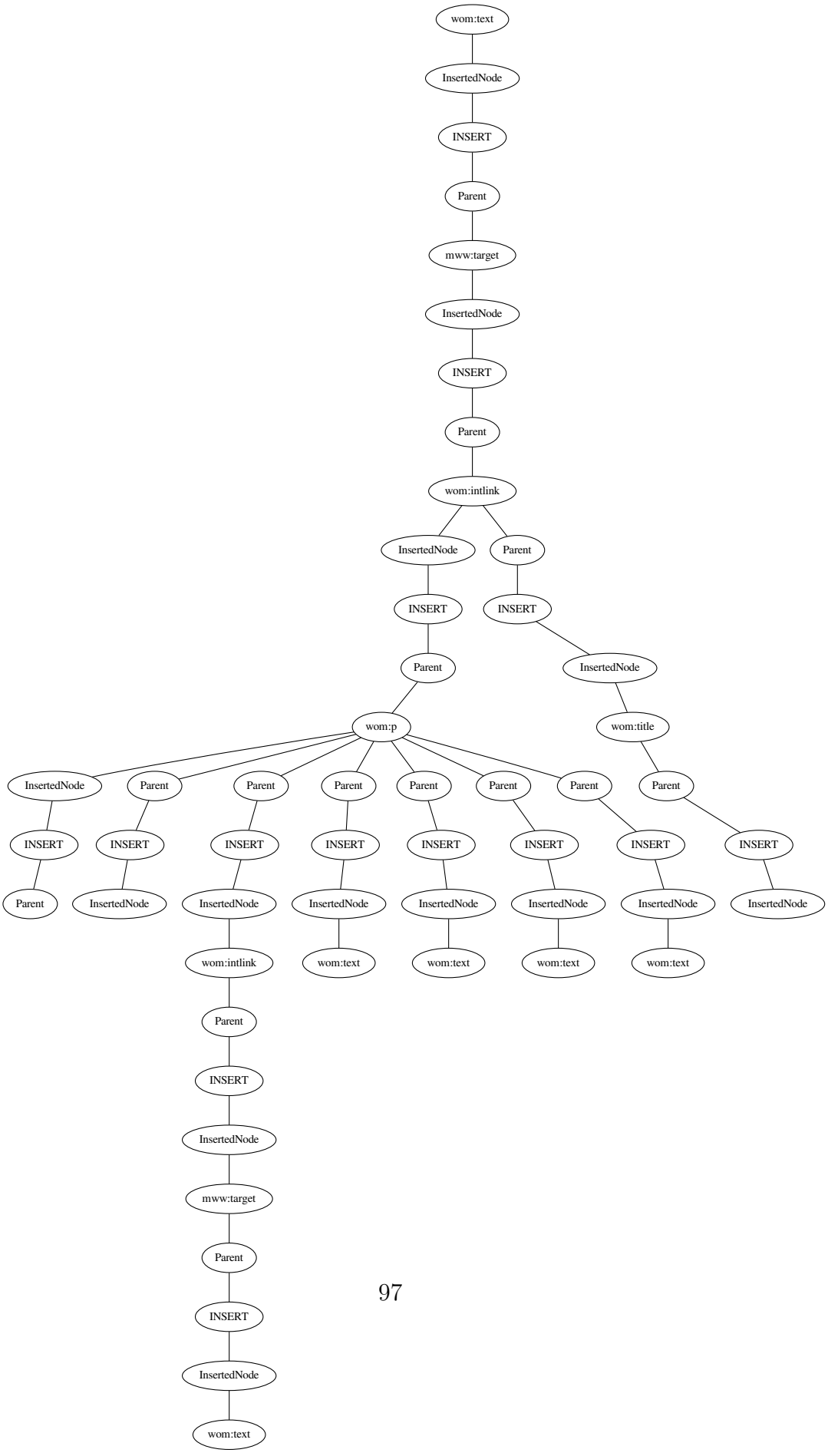
graph has occurrences in (at least) the following revision pairs :	revision A id: 2443498 revision B id: 5131858; revision A id: 5042796 revision B id: 6445130; revision A id: 5072619 revision B id: 5072645; revision A id: 6520278 revision B id: 6548195; revision A id: 12748001 revision B id: 13565511;
graph is subgraph of patterns :	
graph is supergraph of patterns :	6 12 17 20 21 22 23 35 39 40 41 42 46 59 60 67 83 93 135 151 184 198 228 248 289

Pattern No. 143


wom:text
InsertedNode
INSERT
Parent
mww:target
InsertedNode
INSERT
Parent
wom:intlink
InsertedNode
Parent
INSERT
INSERT
Parent
InsertedNode
wom:p
wom:title
InsertedNode
Parent
Parent
Parent
Parent
Parent
Parent
Parent
INSERT
INSERT
INSERT
INSERT
INSERT
INSERT
INSERT
INSERT
Parent
InsertedNode
InsertedNode
InsertedNode
InsertedNode
InsertedNode
InsertedNode
InsertedNode
wom:body
wom:intlink
wom:text
wom:text
wom:text
wom:text
Parent
Parent
INSERT
INSERT
InsertedNode
InsertedNode
wom:text
mww:target
Parent
INSERT
InsertedNode
wom:text

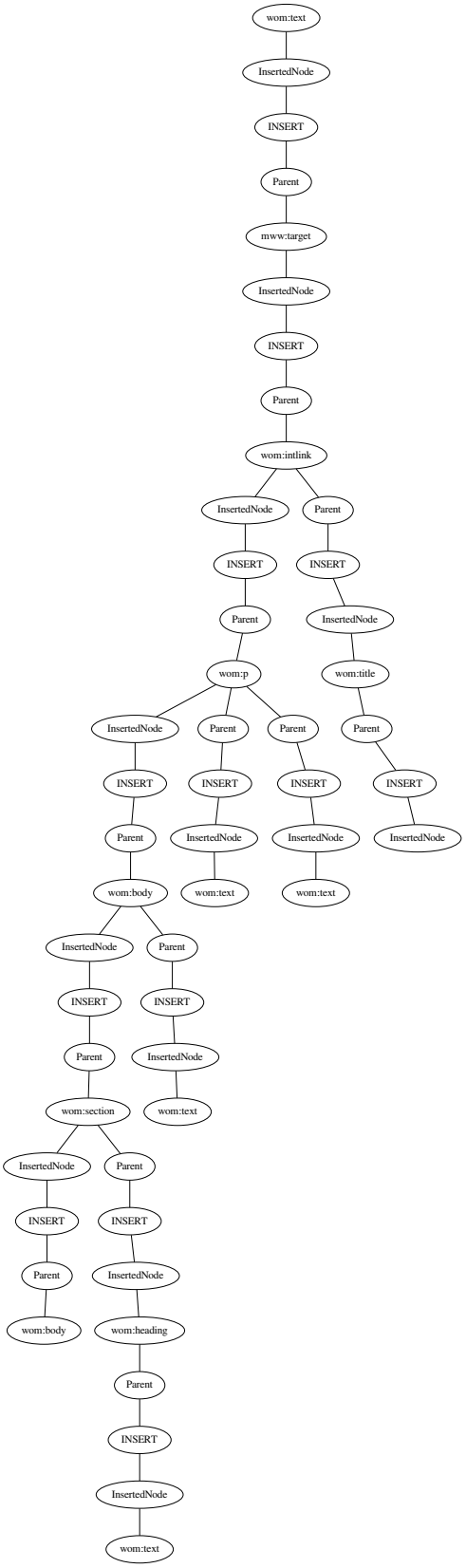
graph has occurrences in (at least) the following revision pairs :	revision A id: 14223369 revision B id: 14223881; revision A id: 246808850 revision B id: 247360688; revision A id: 362265788 revision B id: 363129554; revision A id: 382022859 revision B id: 382032428; revision A id: 382032428 revision B id: 382046870;
graph is subgraph of patterns :	
graph is supergraph of patterns :	3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 21 23 26 34 43 44 51 52 53 54 55 56 57 61 65 66 68 69 81 92 113 114 115 116 120 125 126 166 167 170 173 174 176 178 201 202 207 208 222 237 240 241 253 254 271 274 275 276 281 282 284

Pattern No. 144


wom:text
InsertedNode
INSERT
Parent
mww:target
InsertedNode
INSERT
Parent
wom:intlink
InsertedNode
Parent
INSERT
INSERT
Parent
InsertedNode
wom:p
wom:title
InsertedNode
Parent
Parent
Parent
Parent
Parent
Parent
Parent
INSERT
INSERT
INSERT
INSERT
INSERT
INSERT
INSERT
INSERT
Parent
InsertedNode
InsertedNode
InsertedNode
InsertedNode
InsertedNode
InsertedNode
InsertedNode
wom:intlink
wom:text
wom:text
wom:text
wom:text
Parent
INSERT
InsertedNode
mww:target
Parent
INSERT
InsertedNode
wom:text

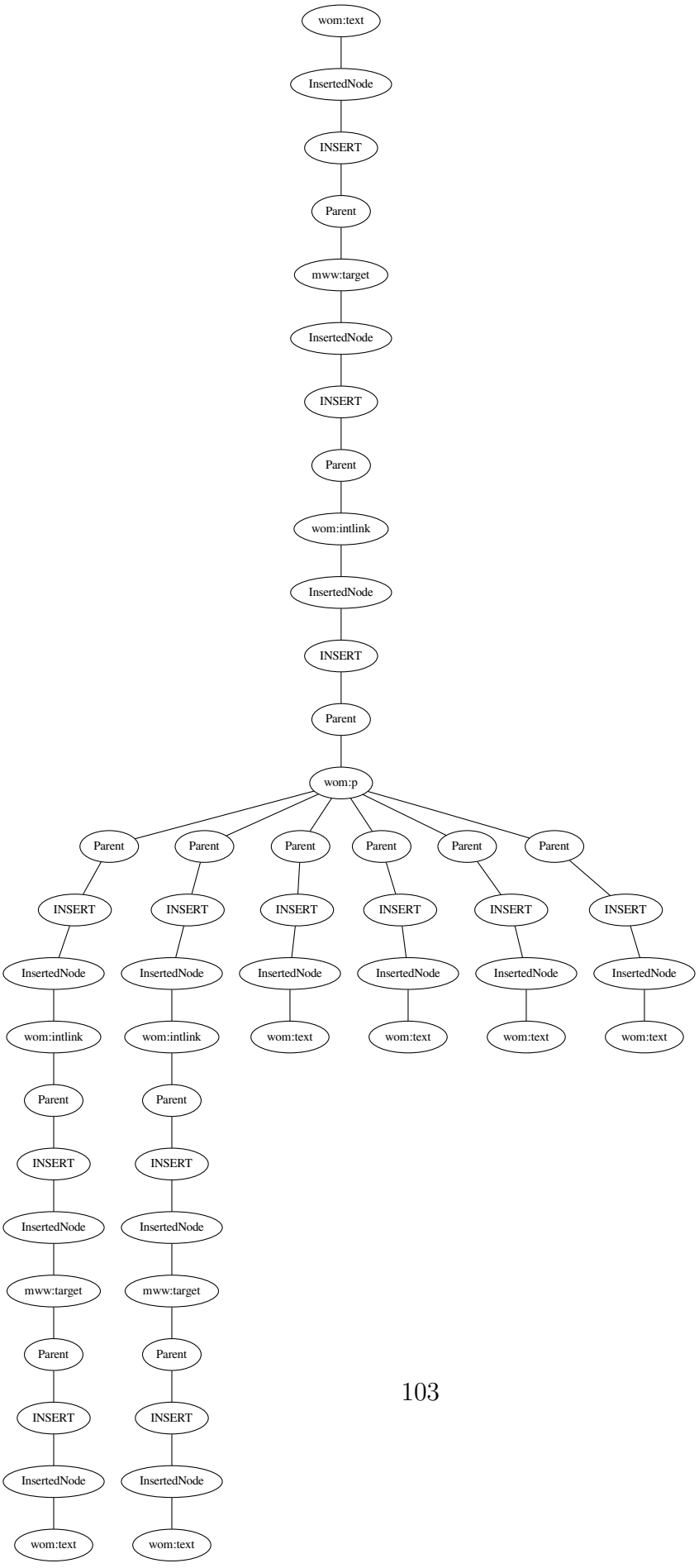
graph has occurrences in (at least) the following revision pairs :	revision A id: 5359016 revision B id: 8141976; revision A id: 14223369 revision B id: 14223881; revision A id: 44286761 revision B id: 46077804; revision A id: 127355053 revision B id: 127403088; revision A id: 208374819 revision B id: 208376565;
graph is subgraph of patterns :	142
graph is supergraph of patterns :	3 4 5 6 7 8 9 10 11 12 13 14 15 17 18 21 23 43 44 51 52 53 54 56 57 65 68 69 81 92 114 115 116 126 166 167 170 178 202 222 237 240 241 253 254 276 281 282 284

Pattern No. 145


wom:text
InsertedNode
INSERT
Parent
mww:target
InsertedNode
INSERT
Parent
wom:intlink
InsertedNode
Parent
INSERT
INSERT
Parent
InsertedNode
wom:p
wom:title
InsertedNode
Parent
Parent
Parent
INSERT
INSERT
INSERT
INSERT
Parent
InsertedNode
InsertedNode
InsertedNode
wom:body
wom:text
wom:text
InsertedNode
Parent
INSERT
INSERT
Parent
InsertedNode
wom:section
wom:text
InsertedNode
Parent
INSERT
INSERT
Parent
InsertedNode
wom:body
wom:heading
Parent
INSERT
InsertedNode
wom:text

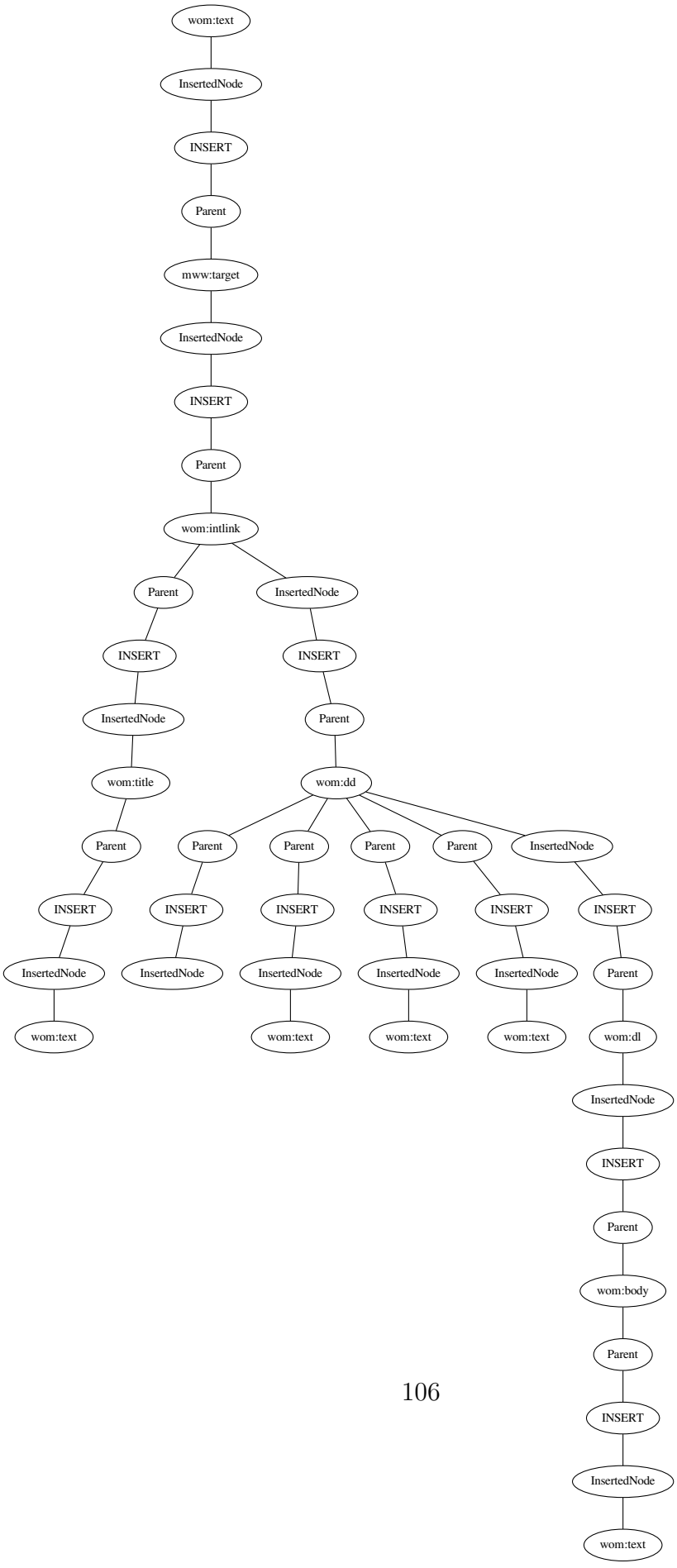
graph has occurrences in (at least) the following revision pairs :	revision A id: 4991103 revision B id: 5000801; revision A id: 5359016 revision B id: 8141976; revision A id: 6965245 revision B id: 6965270; revision A id: 14223369 revision B id: 14223881; revision A id: 16130709 revision B id: 24768319;
graph is subgraph of patterns :	
graph is supergraph of patterns :	3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 21 23 26 34 44 52 55 56 61 66 69 92 101 102 113 114 115 119 120 134 201 207 208 271 275 282 284 286 290

Pattern No. 149


wom:text
InsertedNode
INSERT
Parent
mww:target
InsertedNode
INSERT
Parent
wom:intlink
InsertedNode
INSERT
Parent
wom:p
Parent
Parent
Parent
Parent
Parent
Parent
INSERT
INSERT
INSERT
INSERT
INSERT
INSERT
InsertedNode
InsertedNode
InsertedNode
InsertedNode
InsertedNode
InsertedNode
wom:intlink
wom:intlink
wom:text
wom:text
wom:text
wom:text
Parent
Parent
INSERT
INSERT
InsertedNode
InsertedNode
mww:target
mww:target
Parent
Parent
INSERT
INSERT
InsertedNode
InsertedNode
wom:text
wom:text

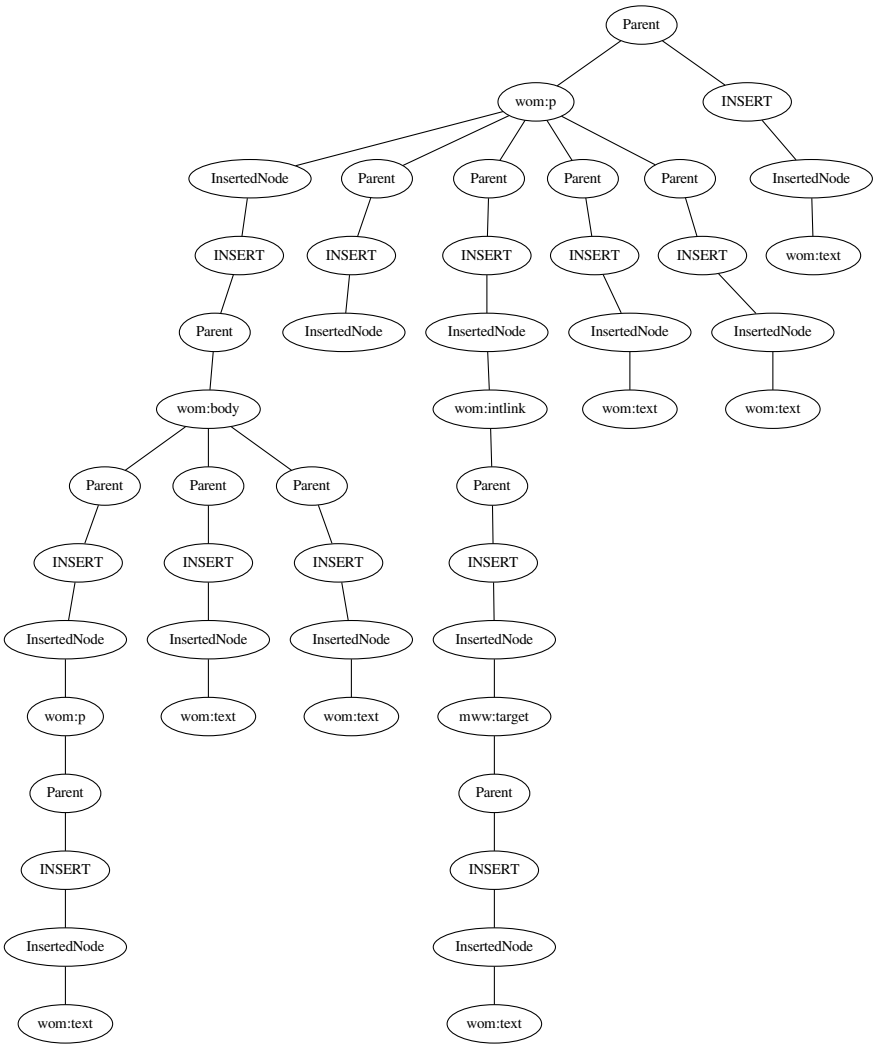
graph has occurrences in (at least) the following revision pairs :	revision A id: 1378554 revision B id: 1378561; revision A id: 4734004 revision B id: 4734726; revision A id: 5359016 revision B id: 8141976; revision A id: 7421939 revision B id: 7720063; revision A id: 16077479 revision B id: 102146373;
graph is subgraph of patterns :	142
graph is supergraph of patterns :	3 4 5 6 7 8 9 10 11 13 14 15 18 43 44 51 54 57 65 68 69 81 88 116 166 167 170 202 222 237 240 253 254 267 276

Pattern No. 164


wom:text
InsertedNode
INSERT
Parent
mww:target
InsertedNode
INSERT
Parent
wom:intlink
Parent
InsertedNode
INSERT
INSERT
InsertedNode
Parent
wom:title
wom:dd
Parent
Parent
Parent
Parent
Parent
InsertedNode
INSERT
INSERT
INSERT
INSERT
INSERT
INSERT
InsertedNode
InsertedNode
InsertedNode
InsertedNode
InsertedNode
Parent
wom:text
wom:text
wom:text
wom:text
wom:dl
InsertedNode
INSERT
Parent
wom:body
Parent
INSERT
InsertedNode
wom:text

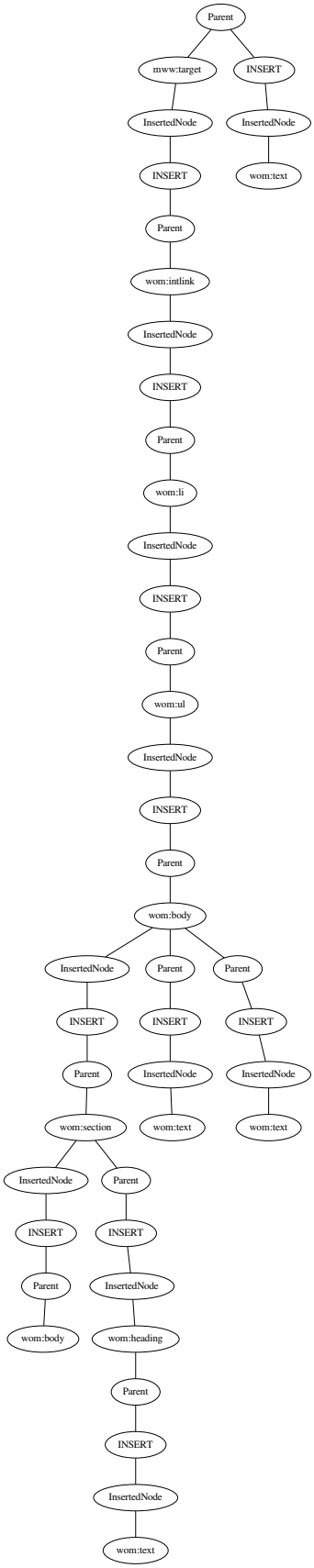
graph has occurrences in (at least) the following revision pairs :	revision A id: 49157044 revision B id: 49157157; revision A id: 57771190 revision B id: 57824325; revision A id: 139221613 revision B id: 139226806; revision A id: 151423544 revision B id: 151424123; revision A id: 179973422 revision B id: 180115908;
graph is subgraph of patterns :	
graph is supergraph of patterns :	6 12 17 20 21 22 23 24 25 27 28 29 30 31 32 33 192 209 215 217 255 258 285 300 301

Pattern No. 175



graph has occurrences in (at least) the following revision pairs :	revision A id: 3034654 revision B id: 3034706; revision A id: 9340385 revision B id: 9361391; revision A id: 58893019 revision B id: 59142580; revision A id: 77074188 revision B id: 77074256; revision A id: 135381416 revision B id: 135384471;
graph is subgraph of patterns :	
graph is supergraph of patterns :	3 4 5 6 7 8 9 10 11 13 14 15 16 18 19 26 34 37 44 52 55 56 57 61 63 64 65 66 68 69 111 113 114 115 116 120 125 126 140 153 154 155 166 167 170 193 201 207 208 237 240 241 242 244 250 251 254 268 270 271 274 275 309 310

Pattern No. 185


Parent
mww:target
INSERT
InsertedNode
InsertedNode
INSERT
wom:text
Parent
wom:intlink
InsertedNode
INSERT
Parent
wom:li
InsertedNode
INSERT
Parent
wom:ul
InsertedNode
INSERT
Parent
wom:body
InsertedNode
Parent
Parent
INSERT
INSERT
INSERT
Parent
InsertedNode
InsertedNode
wom:section
wom:text
wom:text
InsertedNode
Parent
INSERT
INSERT
Parent
InsertedNode
wom:body
wom:heading
Parent
INSERT
InsertedNode
wom:text

graph has occurrences in (at least) the following revision pairs :	revision A id: 6380368 revision B id: 6470369; revision A id: 9172344 revision B id: 9406450; revision A id: 12262527 revision B id: 12263196; revision A id: 12748001 revision B id: 13565511; revision A id: 16316290 revision B id: 25213012;
graph is subgraph of patterns :	
graph is supergraph of patterns :	6 37 60 64 93 101 102 134 139 140 286 288 313

Pattern No. 186

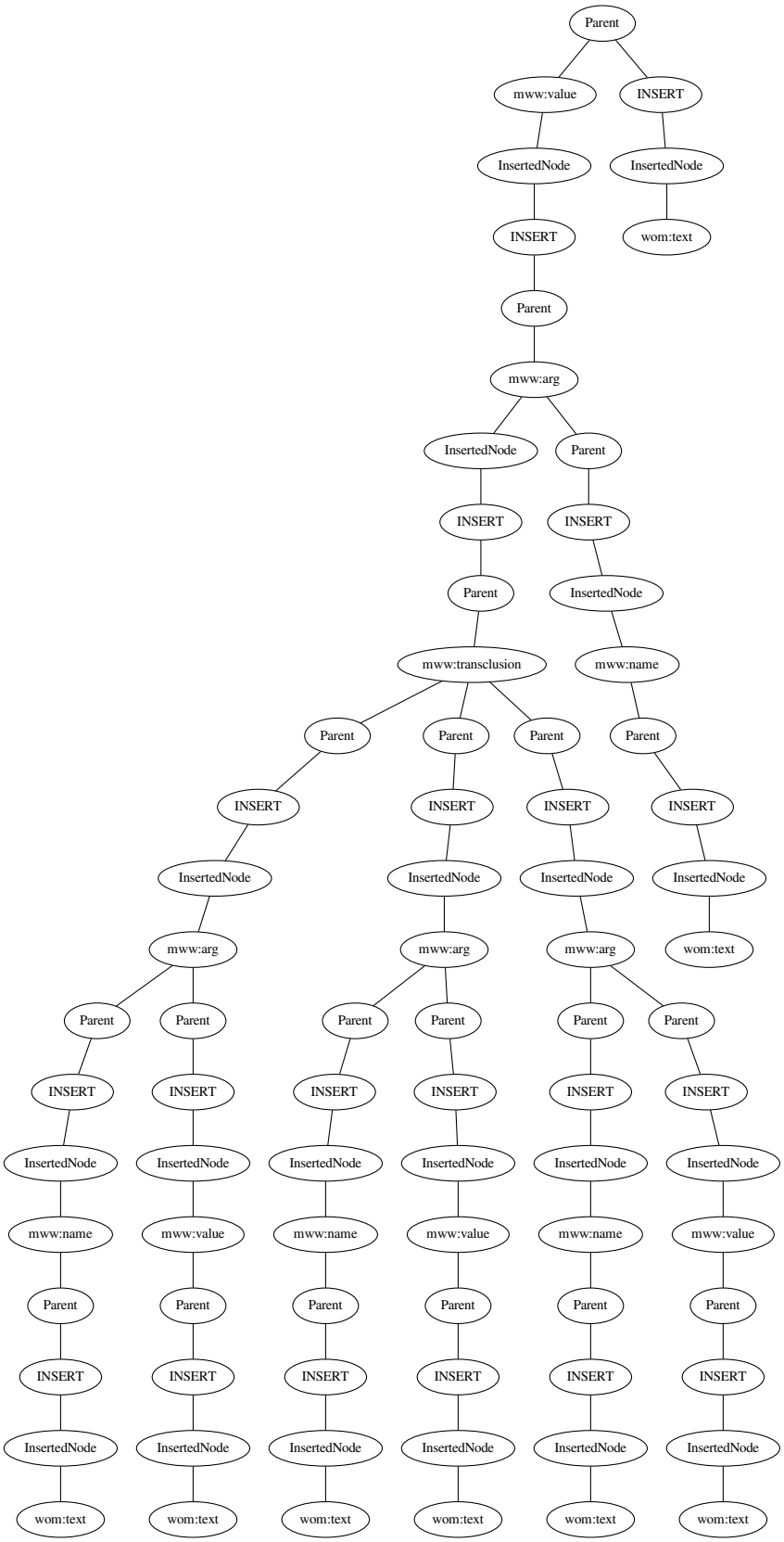
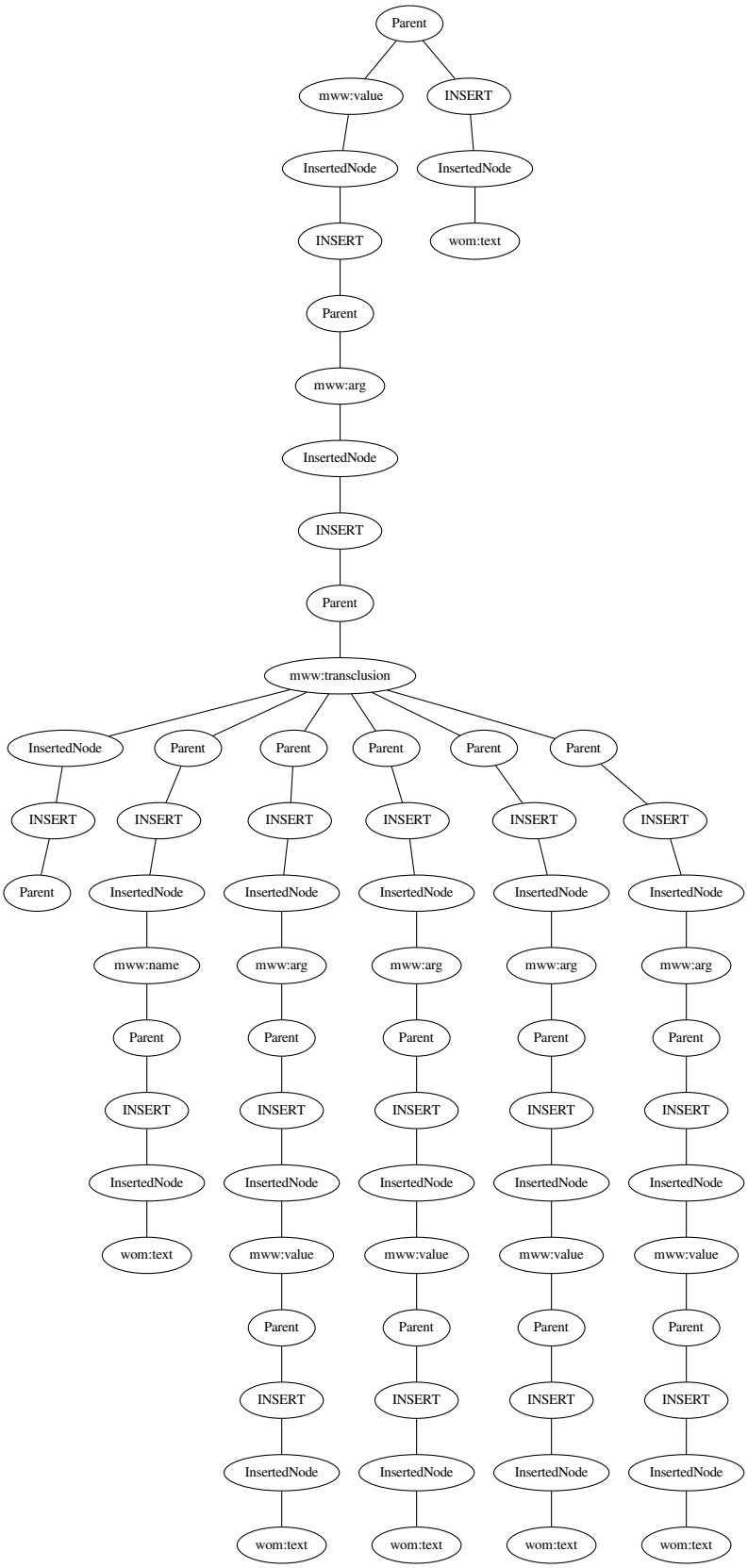


Figure 2.16: Pattern no. 186

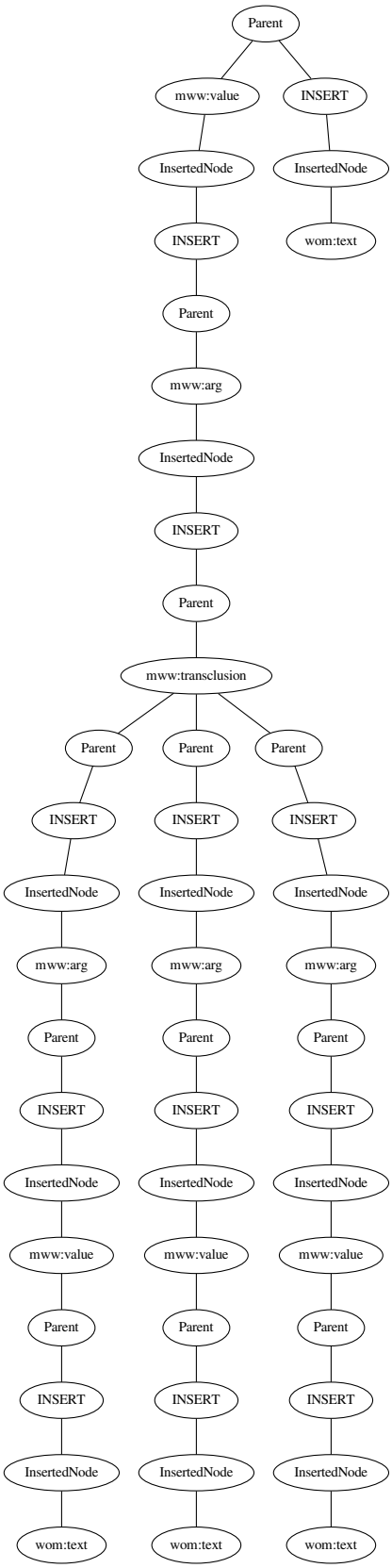
graph has occurrences in (at least) the following revision pairs :	revision A id: 40259015 revision B id: 41645610; revision A id: 72502662 revision B id: 82348989; revision A id: 90853654 revision B id: 106357619; revision A id: 135052356 revision B id: 153450757; revision A id: 145120921 revision B id: 145545842;
graph is subgraph of patterns :	
graph is supergraph of patterns :	2 95 172

Pattern No. 187


Parent
mww:value
INSERT
InsertedNode
InsertedNode
INSERT
wom:text
Parent
mww:arg
InsertedNode
INSERT
Parent
mww:transclusion
InsertedNode
Parent
Parent
Parent
Parent
Parent
INSERT
INSERT
INSERT
INSERT
INSERT
INSERT
Parent
InsertedNode
InsertedNode
InsertedNode
InsertedNode
InsertedNode
mww:name
mww:arg
mww:arg
mww:arg
mww:arg
Parent
Parent
Parent
Parent
Parent
INSERT
INSERT
INSERT
INSERT
INSERT
InsertedNode
InsertedNode
InsertedNode
InsertedNode
InsertedNode
wom:text
mww:value
mww:value
mww:value
mww:value
Parent
Parent
Parent
Parent
INSERT
INSERT
INSERT
INSERT
InsertedNode
InsertedNode
InsertedNode
InsertedNode
wom:text
wom:text
wom:text
wom:text

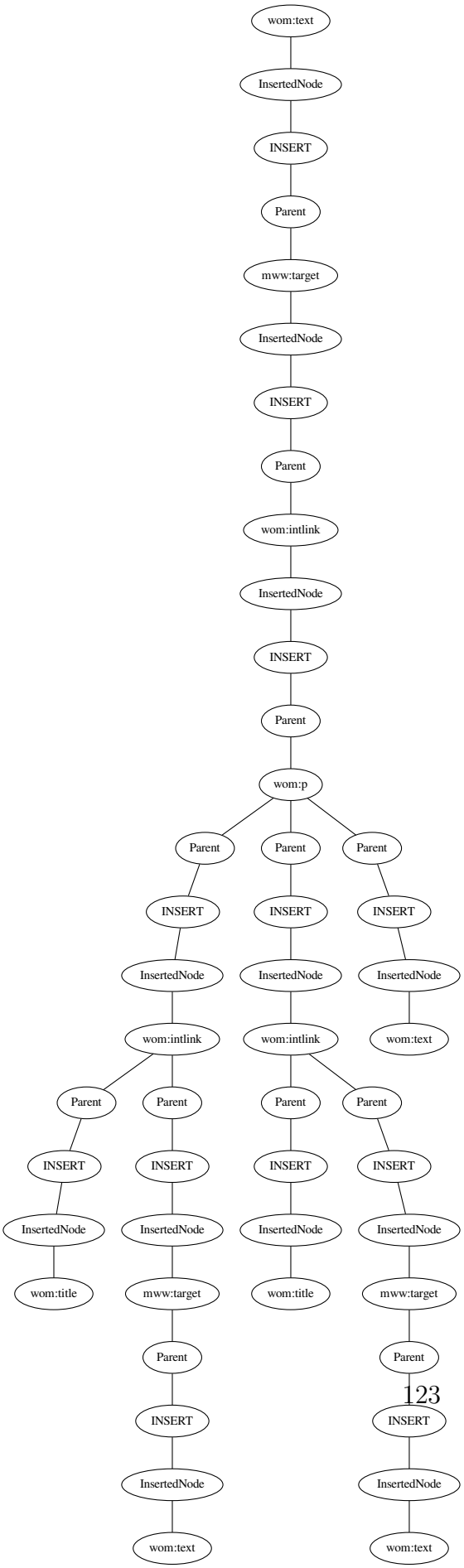
graph has occurrences in (at least) the following revision pairs :	revision A id: 48080311 revision B id: 56106659; revision A id: 78800907 revision B id: 107533037; revision A id: 170664988 revision B id: 170817765; revision A id: 268363914 revision B id: 268364383; revision A id: 343278029 revision B id: 346817266;
graph is subgraph of patterns :	
graph is supergraph of patterns :	1 2 38 84 172

Pattern No. 188


Parent
mww:value
INSERT
InsertedNode
InsertedNode
INSERT
wom:text
Parent
mww:arg
InsertedNode
INSERT
Parent
mww:transclusion
Parent
Parent
Parent
INSERT
INSERT
INSERT
InsertedNode
InsertedNode
InsertedNode
mww:arg
mww:arg
mww:arg
Parent
Parent
Parent
INSERT
INSERT
INSERT
InsertedNode
InsertedNode
InsertedNode
mww:value
mww:value
mww:value
Parent
Parent
Parent
INSERT
INSERT
INSERT
InsertedNode
InsertedNode
InsertedNode
wom:text
wom:text
wom:text

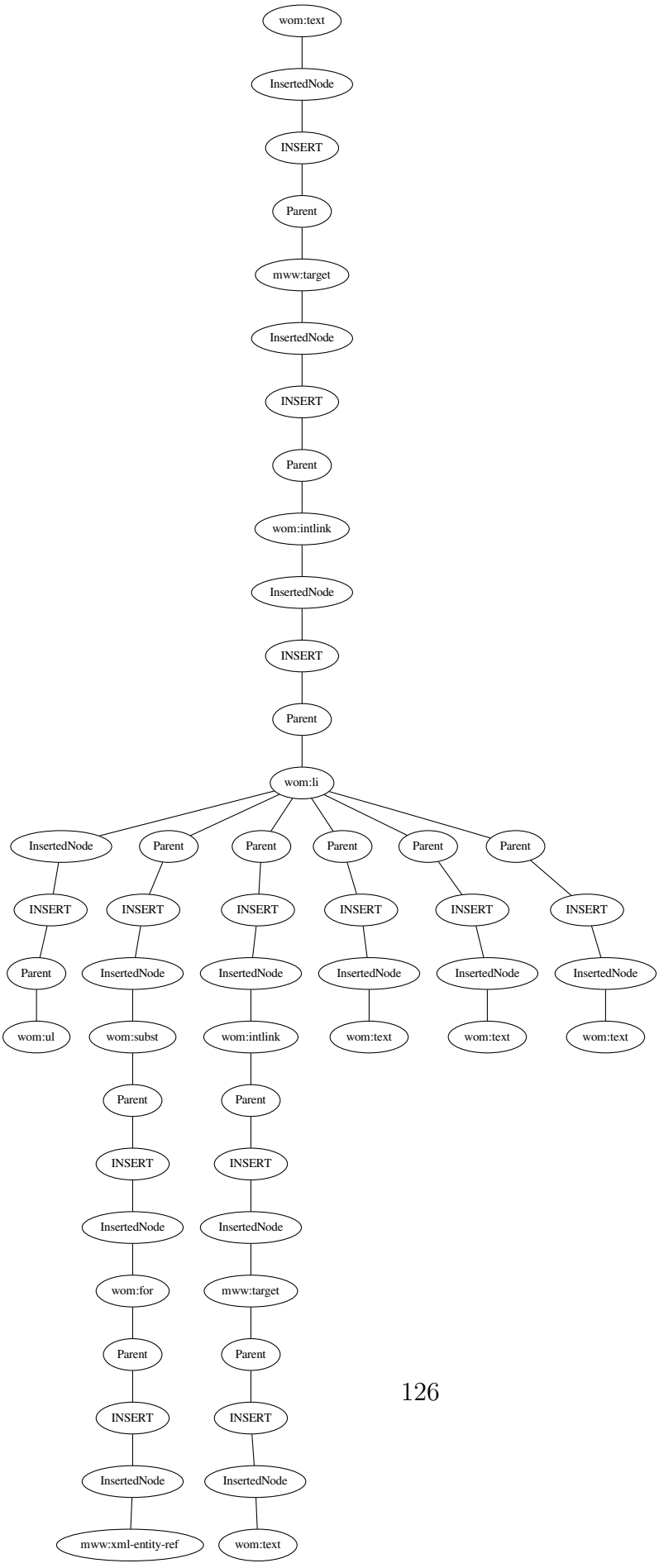
graph has occurrences in (at least) the following revision pairs :	revision A id: 13518044 revision B id: 16942617; revision A id: 40259015 revision B id: 41645610; revision A id: 46586073 revision B id: 55476879; revision A id: 48080311 revision B id: 56106659; revision A id: 51071612 revision B id: 55936807;
graph is subgraph of patterns :	78
graph is supergraph of patterns :	2 172

Pattern No. 220


wom:text
InsertedNode
INSERT
Parent
mww:target
InsertedNode
INSERT
Parent
wom:intlink
InsertedNode
INSERT
Parent
wom:p
Parent
Parent
Parent
INSERT
INSERT
INSERT
InsertedNode
InsertedNode
InsertedNode
wom:intlink
wom:intlink
wom:text
Parent
Parent
Parent
Parent
INSERT
INSERT
INSERT
INSERT
InsertedNode
InsertedNode
InsertedNode
InsertedNode
wom:title
mww:target
wom:title
mww:target
Parent
Parent
INSERT
INSERT
InsertedNode
InsertedNode
wom:text
wom:text

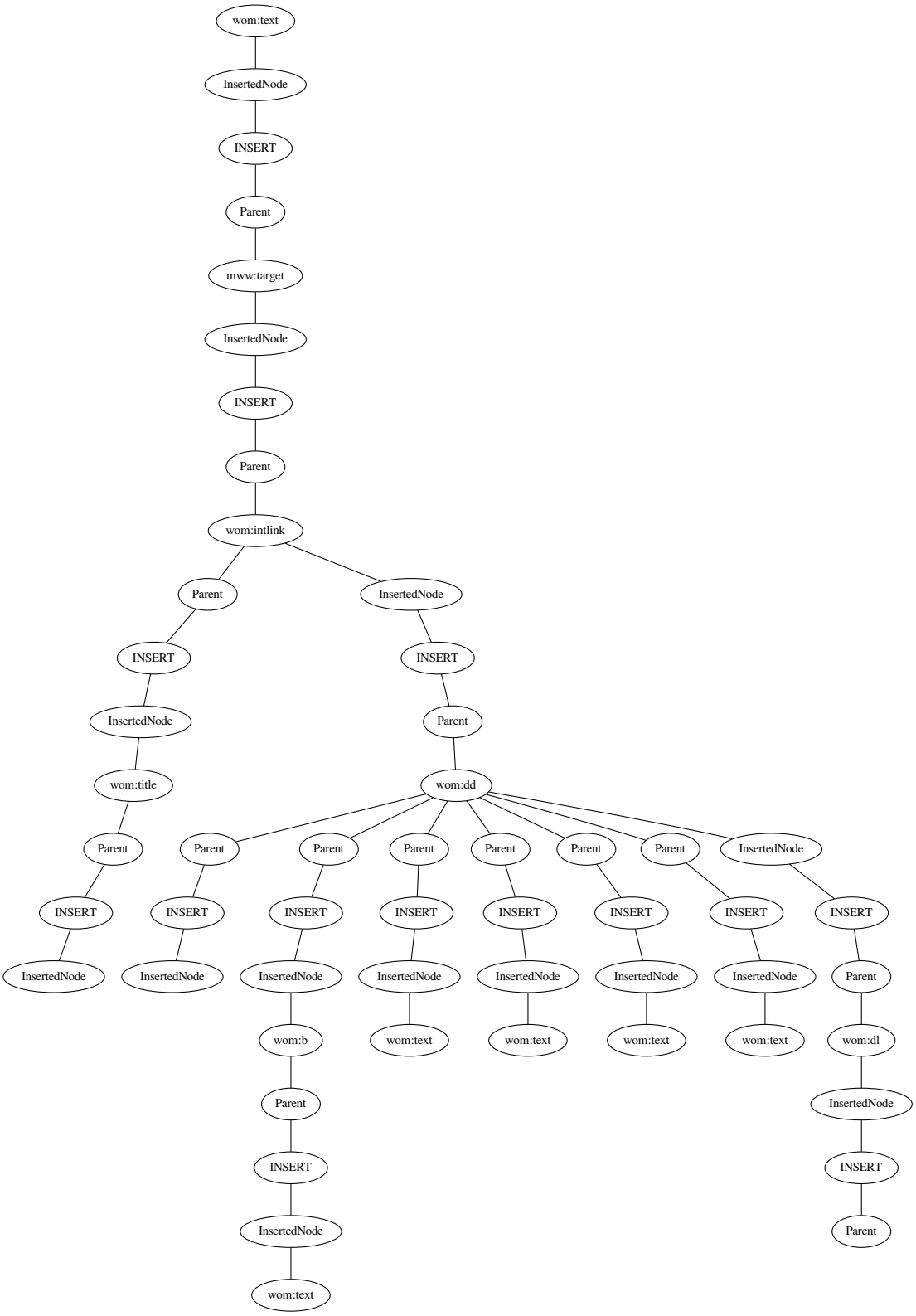
graph has occurrences in (at least) the following revision pairs :	revision A id: 175239 revision B id: 175260; revision A id: 621714 revision B id: 622105; revision A id: 1207572 revision B id: 1238684; revision A id: 5359016 revision B id: 8141976; revision A id: 15595117 revision B id: 15595255;
graph is subgraph of patterns :	142 278
graph is supergraph of patterns :	3 4 5 6 7 8 9 10 11 12 14 15 17 18 54 65 68 92 167 222 254 281 282 284

Pattern No. 235


wom:text
InsertedNode
INSERT
Parent
mww:target
InsertedNode
INSERT
Parent
wom:intlink
InsertedNode
INSERT
Parent
wom:li
InsertedNode
Parent
Parent
Parent
Parent
Parent
INSERT
INSERT
INSERT
INSERT
INSERT
INSERT
Parent
InsertedNode
InsertedNode
InsertedNode
InsertedNode
InsertedNode
wom:ul
wom:subst
wom:intlink
wom:text
wom:text
wom:text
Parent
Parent
INSERT
INSERT
InsertedNode
InsertedNode
wom:for
mww:target
Parent
Parent
INSERT
INSERT
InsertedNode
InsertedNode
mww:xml-entity-ref
wom:text

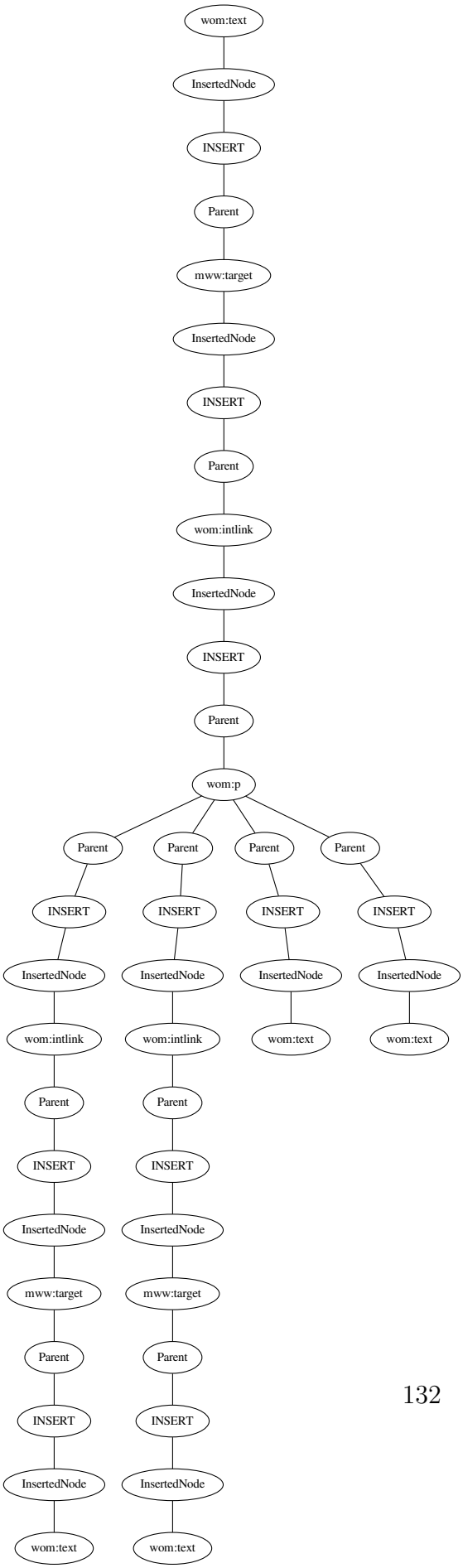
graph has occurrences in (at least) the following revision pairs :	revision A id: 290011761 revision B id: 291655560; revision A id: 325813725 revision B id: 326126815; revision A id: 343339961 revision B id: 343564464; revision A id: 346376733 revision B id: 346378833; revision A id: 361997002 revision B id: 363448256;
graph is subgraph of patterns :	
graph is supergraph of patterns :	6 35 39 40 42 46 58 59 60 67 82 83 93 135 151 197 198 236 243 247 289

Pattern No. 246


wom:text
InsertedNode
INSERT
Parent
mww:target
InsertedNode
INSERT
Parent
wom:intlink
Parent
InsertedNode
INSERT
INSERT
InsertedNode
Parent
wom:title
wom:dd
Parent
Parent
Parent
Parent
Parent
Parent
Parent
InsertedNode
INSERT
INSERT
INSERT
INSERT
INSERT
INSERT
INSERT
INSERT
InsertedNode
InsertedNode
InsertedNode
InsertedNode
InsertedNode
InsertedNode
InsertedNode
Parent
wom:b
wom:text
wom:text
wom:text
wom:text
wom:dl
Parent
InsertedNode
INSERT
InsertedNode
Parent
wom:text

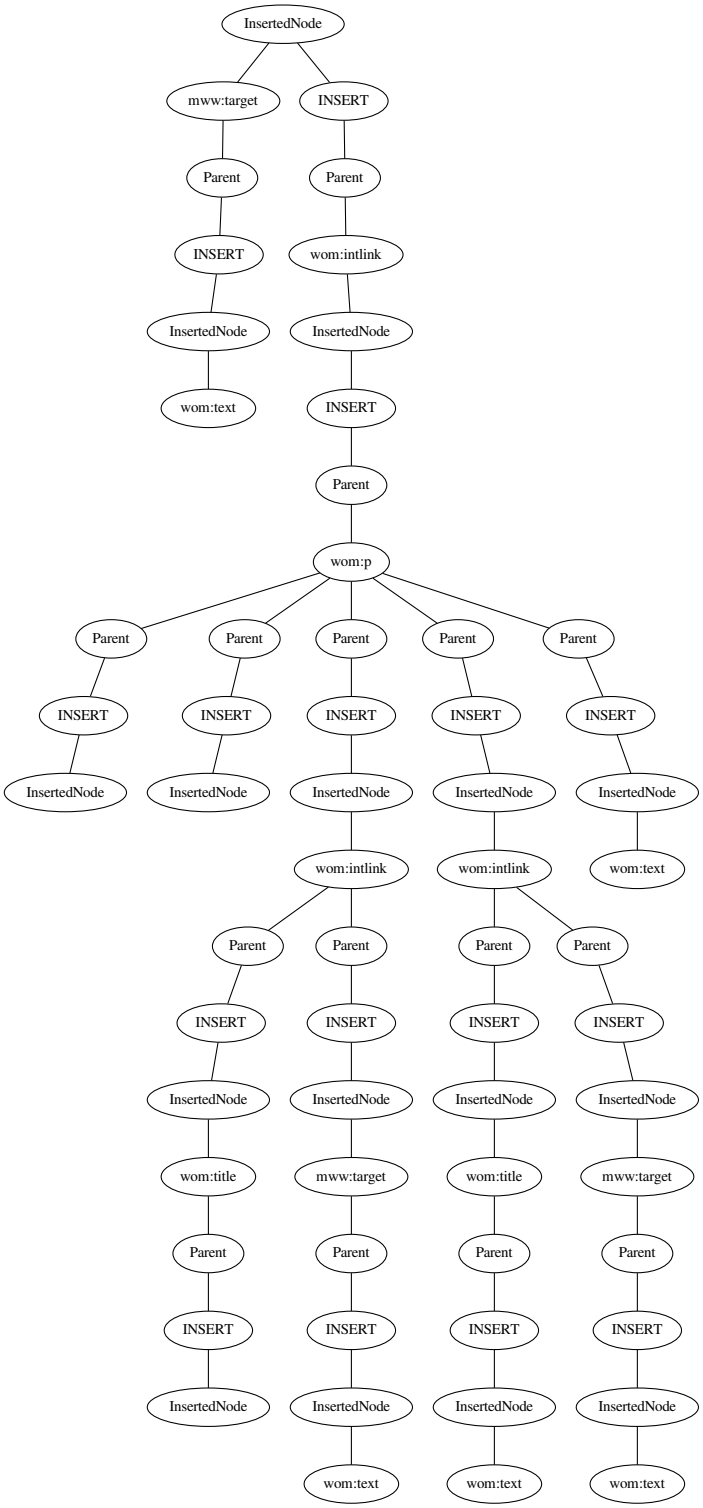
graph has occurrences in (at least) the following revision pairs :	revision A id: 14797615 revision B id: 14798837; revision A id: 375703076 revision B id: 375718764; revision A id: 380561896 revision B id: 380563345;
graph is subgraph of patterns :	
graph is supergraph of patterns :	6 12 17 21 23 24 25 27 28 29 30 31 32 33 209 215 217

Pattern No. 267


wom:text
InsertedNode
INSERT
Parent
mww:target
InsertedNode
INSERT
Parent
wom:intlink
InsertedNode
INSERT
Parent
wom:p
Parent
Parent
Parent
Parent
INSERT
INSERT
INSERT
INSERT
InsertedNode
InsertedNode
InsertedNode
InsertedNode
wom:intlink
wom:intlink
wom:text
wom:text
Parent
Parent
INSERT
INSERT
InsertedNode
InsertedNode
mww:target
mww:target
Parent
Parent
INSERT
INSERT
InsertedNode
InsertedNode
wom:text
wom:text

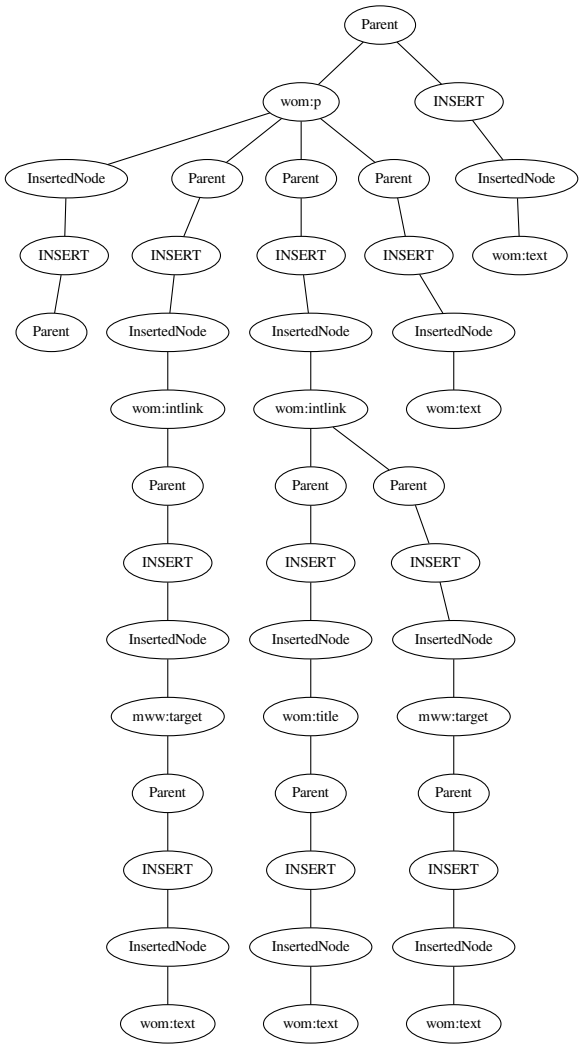
graph has occurrences in (at least) the following revision pairs :	revision A id: 1378554 revision B id: 1378561; revision A id: 1464792 revision B id: 2754088; revision A id: 4734004 revision B id: 4734726; revision A id: 5359016 revision B id: 8141976; revision A id: 7331750 revision B id: 7331807;
graph is subgraph of patterns :	88 142 148 149
graph is supergraph of patterns :	3 4 5 6 7 8 9 10 11 13 14 15 18 44 51 54 57 65 68 69 116 167 222 240 254

Pattern No. 278



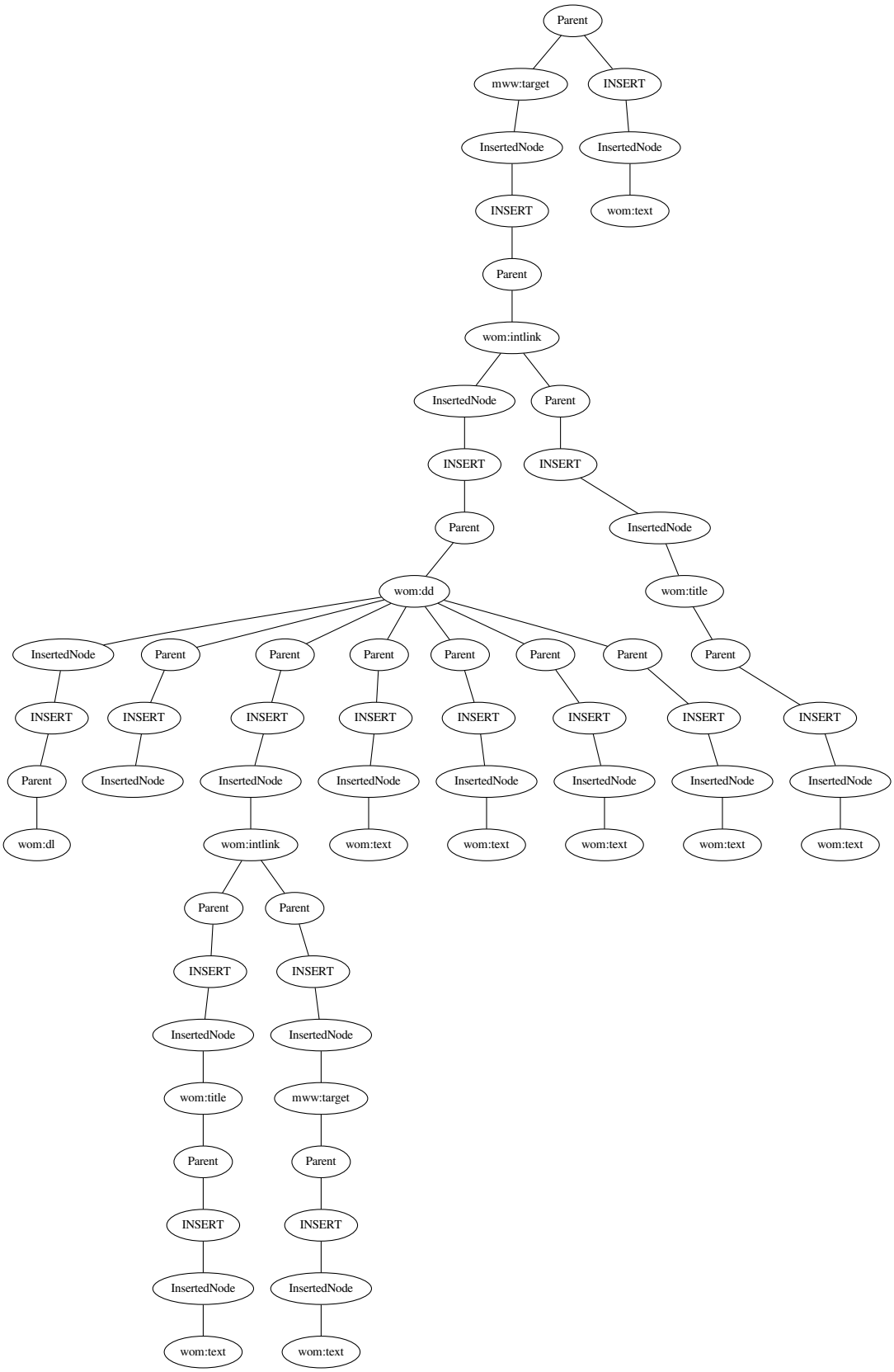
graph has occurrences in (at least) the following revision pairs :	revision A id: 1207572 revision B id: 1238684; revision A id: 5359016 revision B id: 8141976; revision A id: 103155489 revision B id: 110087167; revision A id: 216324329 revision B id: 216436425; revision A id: 244588905 revision B id: 271529496;
graph is subgraph of patterns :	142
graph is supergraph of patterns :	3 4 5 6 7 8 9 10 11 12 14 15 17 18 20 21 22 23 50 54 65 68 79 91 92 137 167 220 222 253 254 279 281 282 284

Pattern No. 297



graph has occurrences in (at least) the following revision pairs :	revision A id: 308634 revision B id: 308661; revision A id: 621714 revision B id: 622105; revision A id: 1464792 revision B id: 2754088; revision A id: 3034654 revision B id: 3034706; revision A id: 5359016 revision B id: 8141976;
graph is subgraph of patterns :	121 142
graph is supergraph of patterns :	3 4 5 6 7 8 9 10 11 12 13 14 15 17 18 20 21 22 23 44 50 51 52 53 54 56 57 65 68 69 90 91 92 114 115 116 122 137 147 155 167 177 178 222 241 254 272 279 281 282 284

Pattern No. 299


Parent
mww:target
INSERT
InsertedNode
InsertedNode
INSERT
wom:text
Parent
wom:intlink
InsertedNode
Parent
INSERT
INSERT
Parent
InsertedNode
wom:dd
wom:title
InsertedNode
Parent
Parent
Parent
Parent
Parent
Parent
Parent
INSERT
INSERT
INSERT
INSERT
INSERT
INSERT
INSERT
INSERT
Parent
InsertedNode
InsertedNode
InsertedNode
InsertedNode
InsertedNode
InsertedNode
InsertedNode
wom:dl
wom:intlink
wom:text
wom:text
wom:text
wom:text
wom:text
Parent
Parent
INSERT
INSERT
InsertedNode
InsertedNode
wom:title
mww:target
Parent
Parent
INSERT
INSERT
InsertedNode
InsertedNode
wom:text
wom:text

graph has occurrences in (at least) the following revision pairs :	revision A id: 3489847 revision B id: 4061708; revision A id: 125522905 revision B id: 125522953; revision A id: 139221613 revision B id: 139226806; revision A id: 204869581 revision B id: 204871924; revision A id: 211901416 revision B id: 211901998;
graph is subgraph of patterns :	
graph is supergraph of patterns :	6 12 17 20 21 22 23 24 25 28 29 31 192 209 217 258

References

Abouzeid, A., Bajda-Pawlikowski, K., Abadi, D., Silberschatz, A., & Rasin, A. (2009, August). Hadoopdb: an architectural hybrid of mapreduce and dbms technologies for analytical workloads. *Proc. VLDB Endow. 2*(1), 922–933. doi:10.14778/1687627.1687731

Aggarwal, C. C. (2015). *Data mining - the textbook*. Springer. doi:10.1007/978-3-319-14142-8

Aggarwal, C. C. & Han, J. (Eds.). (2014). *Frequent pattern mining*. Springer. doi:10.1007/978-3-319-07821-2

Cormen, T. H., Stein, C., Rivest, R. L., & Leiserson, C. E. (2001). *Introduction to algorithms* (2nd). McGraw-Hill Higher Education.

Dean, J. & Ghemawat, S. (2008, January). Mapreduce: simplified data processing on large clusters. *Commun. ACM*, *51*(1), 107–113. doi:10.1145/1327452.1327492

Dohrn, H. & Riehle, D. (2011). Design and implementation of the sweble wikitext parser: unlocking the structured data of wikipedia. In *Proceedings of the 7th international symposium on wikis and open collaboration* (pp. 72–81). WikiSym '11. Mountain View, California: ACM.

Dohrn, H. & Riehle, D. (2013). Design and implementation of wiki content transformations and refactorings. In *Proceedings of the 9th international symposium on open collaboration* (2:1–2:10). WikiSym '13. Hong Kong, China: ACM.

Dohrn, H. & Riehle, D. (2014). Fine-grained change detection in structured text documents. In *Proceedings of the 2014 acm symposium on document engineering* (pp. 87–96). DocEng '14. Fort Collins, Colorado, USA: ACM.

Ferrucci, D. A. (2011, June). Ibm's watson/deepqa. *SIGARCH Comput. Archit. News*, *39*(3). doi:10.1145/2024723.2019525

Gärtner, T., Flach, P., & Wrobel, S. (2003). On graph kernels: hardness results and efficient alternatives. In *In: conference on learning theory* (pp. 129–143).

Han, J. (2005). *Data mining: concepts and techniques*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Ketkar, N. S., Holder, L. B., & Cook, D. J. (2005). Subdue: compression-based frequent pattern discovery in graph data. In *Proceedings of the 1st international workshop on open source data mining: frequent pattern mining implementations* (pp. 71–76). OSDM '05. Chicago, Illinois: ACM. doi:10.1145/1133905.1133915

Langdon, M. (2014). *The work of art in a digital age: art, technology and globalisation*. Springer. doi:10.1007/978-1-4939-1270-4

Meinl, T., Wörlein, M., Fischer, I., & Philippsen, M. (2006). Mining molecular datasets on symmetric multiprocessor systems. In *International conference on systems, man and cybernetics, 2006. smc '06* (pp. 1269–1274).

Nijssen, S. & Kok, J. N. (2004). A quickstart in frequent structure mining can make a difference. In *Proceedings of the tenth acm sigkdd international conference on knowledge discovery and data mining* (pp. 647–652). KDD '04. Seattle, WA, USA: ACM. doi:10.1145/1014052.1014134

ParSeMiS project. (n.d.). Retrieved April 20, 2016, from https://github.com/timtadh/parsemis

ParSeMiS - the Parallel and Sequential Mining Suite. (n.d.). Retrieved April 20, 2016, from www2.cs.fau.de/EN/research/zold/ParSeMiS/index.html

Wikipedia statistics. (n.d.). Retrieved April 20, 2016, from https://en.wikipedia.org/wiki/Wikipedia:Statistics

Samatova, N. F., Hendrix, W., Jenkins, J., Padmanabhan, K., & Chakraborty, A. (2013). *Practical graph mining with r*. Chapman & Hall/CRC.

Viégas, F. B., Wattenberg, M., & Dave, K. (2004). Studying cooperation and conflict between authors with history flow visualizations. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 575–582). CHI '04. Vienna, Austria: ACM. doi:10.1145/985692.985765

Yan, X. & Han, J. (2002). Gspan: graph-based substructure pattern mining. In *Proceedings of the 2002 ieee international conference on data mining* (pp. 721–). ICDM '02. Washington, DC, USA: IEEE Computer Society.

Yan, X. & Han, J. (2003). Closegraph: mining closed frequent graph patterns. In *Proceedings of the ninth acm sigkdd international conference on knowledge discovery and data mining* (pp. 286–295). KDD '03. Washington, D.C.: ACM.

Zaki, M. J. (2005, March). Efficiently mining frequent embedded unordered trees. *Fundamenta Informaticae*, *66*(1-2), 33–52. special issue on Advances in Mining Graphs, Trees and Sequences.