The Wikipedia XML Corpus

Ludovic Denoyer and Patrick Gallinari

Laboratoire d'Informatique de Paris 6
8 rue du capitaine Scott
75015 Paris

{ludovic.denoyer,patrick.gallinari}@lip6.fr
http://www-connex.lip6.fr/denoyer/wikipediaXML

Abstract. This article presents the general Wikipedia XML Collection developped for Structured Information Retrieval and Structured Machine Learning. This collection has been built from the Wikipedia Enclyclopedia. We detail particularly here which parts of this collection have been used during INEX 2006 for the Ad-hoc track and for the XML Mining track. Note that other tracks of INEX - multimedia track for example - have also been based on this collection.

1 Introduction

Wikipedia[1] is a well known free content, multilingual encyclopedia written collaboratively by contributors around the world. Anybody can edit an article using a wiki markup language that offers a simplified alternative to HTML. This encyclopedia is composed of millions of articles in different languages.

Content-oriented XML retrieval is an area of Information Retrieval (IR) research that is receiving an increasing interest. There already exists a very active community in the IR/ XML domain which started to work on XML search engines and XML textual data. This community is mainly organized since 2002 around the INEX initiative (INitiative for the Evaluation of XML Retrieval) which is funded by the DELOS network of excellence on Digital Libraries.

In this article, we describe a set of XML collections based on Wikipedia. These collections can be used in a large variety of XML IR/Machine Learning tasks like ad-hoc retrieval, categorization, clustering or structure mapping. These corpora are currently used for both, INEX 2006[2] and the XML Document Mining Challenge[3]. The article provides a description of the corpus.

The collections are downloadable on the website:

- *http://www-connex.lip6.fr/∼denoyer/wikipediaXML*

[1] http://www.wikipedia.org

[2] http://inex.is.informatik.uni-duisburg.de/2006

[3] http://xmlmining.lip6.fr

N. Fuhr, M. Lalmas, and A. Trotman (Eds.): INEX 2006, LNCS 4518, pp. 12–19, 2007.

2 Description of the Corpus

The corpus is composed of 8 main collections corresponding to 8 different languages[4] : English, French, German, Dutch, Spanish, Chinese, Arabian and Japanese. Each collection is a set of XML documents built using Wikipedia and encoded in UTF-8. In addition to these 8 collections, we also provide different *additional collections* for other IR/Machine Learning tasks like categorization and clustering, NLP, machine translation, multimedia IR, entity search, etc.

2.1 Main Collections

The main collections are a set of XML files in 8 different languages. The table 1 gives a detailed description of each collection.

Table 1. General statistics about the *Main Collections*

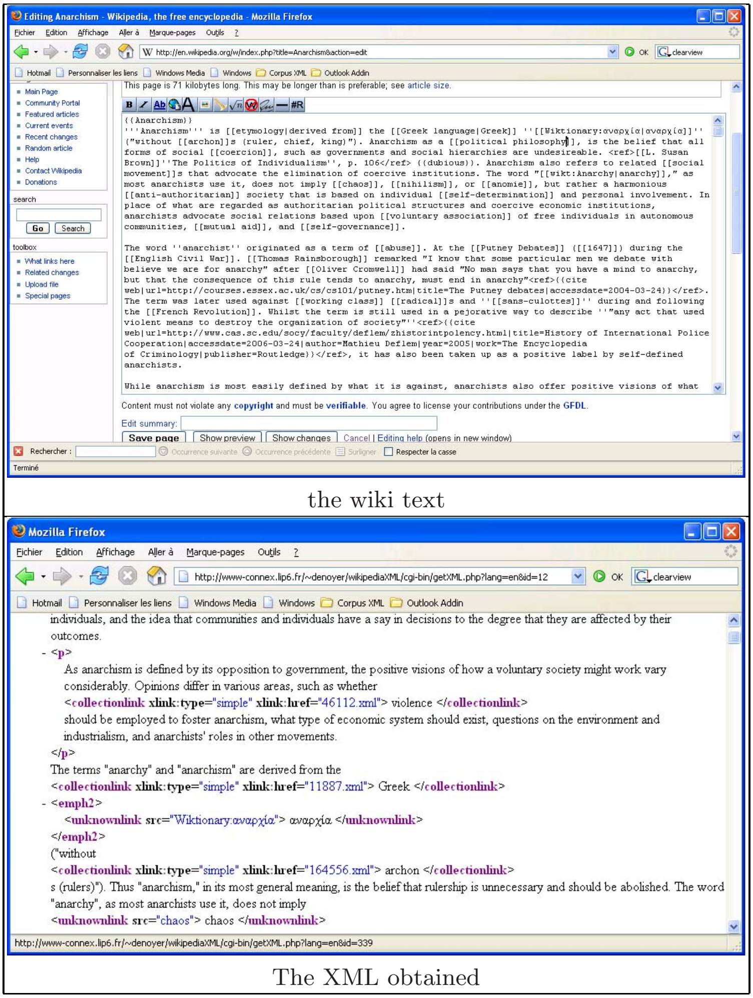
Collection name	Language	Number of documents	Size of the collection (MegaBytes)
main-english	English	659,388	≈ 4,600
20060130_french	French	110,838	≈ 730
20060123_german	German	305,099	≈ 2,079
20060227_dutch	Dutch	125,004	≈ 607
20060130_spanish	Spanish	79,236	≈ 504
20060303_chinese	Chinese	56,661	≈ 360
20060326_arabian	Arabian	11,637	≈ 53
20060303_japanese	Japanese	187,492	≈ 1,425

Each collection contains a set of documents where each filename is a number corresponding to the id of the file (for example : *15243.xml*). Each id is unique and each file corresponds to an article of Wikipedia. We only kept articles and removed all the wikipedia pages corresponding to ”‘Talks”’, ”‘Template”’, etc.. Each file is an UTF-8 document which is created from the wikitext of the original article. Figure 1 gives an example of an English article extracted from the corpus.

Tag labels. We introduced different tags in order to represent the different parts of a document. We distinguish two types of tags:

- The general tags (*article,section, paragraph,etc.*) that do not depend on the language of the collection. These tags correspond to the structural information contained in the wikitext format (for example : == *Main part* == is transformed into <*title*>*Main part*< /*title*>)
- The template tags (*template_infobox,etc.*) represent the information contained into the wikipedia templates. Wikipedia templates are used to represent a repetitive type of information. For example, each country described into wikipedia starts with a table containing its population, language, size,etc.

[4] Some additional languages will be added during the next months.



Editing Anarchism - Wikipedia, the free encyclopedia - Mozilla Firefox
Fichier Edition Affichage Aller à Marque-pages Outils ?
http://en.wikipedia.org/w/index.php?title=Anarchism&action=edit
OK clearview
Hotmail Personnaliser les liens Windows Media Windows Corpus XML Outlook Addin
Main Page
Community Portal
Featured articles
Current events
Recent changes
Random article
Help
Contact Wikipedia
Donations
search
Go Search
toolbox
What links here
Related changes
Upload file
Special pages
Edit summary:
Save page Show preview Show changes Cancel | Editing help (opens in new window)
Rechercher : Occurrence suivante Occurrence précédente Surligner Respecter la casse
Terminé

This page is 71 kilobytes long. This may be longer than is preferable; see article size.

```
{{Anarchism}}
'''Anarchism''' is [[etymology|derived from]] the [[Greek language|Greek]] ''[[Wiktionary:αναρχία|αναρχία]]''
("without [[archon]]s (ruler, chief, king)"). Anarchism as a [[political philosophy]], is the belief that all
forms of social [[coercion]], such as governments and social hierarchies are undesireable. <ref>[[L. Susan
Brown]]''The Politics of Individualism'', p. 106</ref> {{dubious}}. Anarchism also refers to related [[social
movement]]s that advocate the elimination of coercive institutions. The word "[[wikt:Anarchy|anarchy]]," as
most anarchists use it, does not imply [[chaos]], [[nihilism]], or [[anomie]], but rather a harmonious
[[anti-authoritarian]] society that is based on individual [[self-determination]] and personal involvement. In
place of what are regarded as authoritarian political structures and coercive economic institutions,
anarchists advocate social relations based upon [[voluntary association]] of free individuals in autonomous
communities, [[mutual aid]], and [[self-governance]].

The word ''anarchist'' originated as a term of [[abuse]]. At the [[Putney Debates]] ([[1647]]) during the
[[English Civil War]]. [[Thomas Rainsborough]] remarked "I know that some particular men we debate with
believe we are for anarchy" after [[Oliver Cromwell]] had said "No man says that you have a mind to anarchy,
but that the consequence of this rule tends to anarchy, must end in anarchy"<ref>{{cite
web|url=http://courses.essex.ac.uk/cs/cs101/putney.htm|title=The Putney debates|accessdate=2004-03-24}}</ref>.
The term was later used against [[working class]] [[radical]]s and ''[[sans-culottes]]'' during and following
the [[French Revolution]]. Whilst the term is still used in a pejorative way to describe ''"any act that used
violent means to destroy the organization of society"''<ref>{{cite
web|url=http://www.cas.sc.edu/socy/faculty/deflem/zhistorintpolency.html|title=History of International Police
Cooperation|accessdate=2006-03-24|author=Mathieu Deflem|year=2005|work=The Encyclopedia
of Criminology|publisher=Routledge}}</ref>, it has also been taken up as a positive label by self-defined
anarchists.

While anarchism is most easily defined by what it is against, anarchists also offer positive visions of what
```

Content must not violate any copyright and must be verifiable. You agree to license your contributions under the GFDL.

the wiki text

Mozilla Firefox
Fichier Edition Affichage Aller à Marque-pages Outils ?
http://www-connex.lip6.fr/~denoyer/wikipediaXML/cgi-bin/getXML.php?lang=en&id=12
OK clearview
Hotmail Personnaliser les liens Windows Media Windows Corpus XML Outlook Addin
http://www-connex.lip6.fr/~denoyer/wikipediaXML/cgi-bin/getXML.php?lang=en&id=339

```
  individuals, and the idea that communities and individuals have a say in decisions to the degree that they are affected by their
  outcomes.
- <p>
    As anarchism is defined by its opposition to government, the positive visions of how a voluntary society might work vary
    considerably. Opinions differ in various areas, such as whether
    <collectionlink xlink:type="simple" xlink:href="46112.xml"> violence </collectionlink>
    should be employed to foster anarchism, what type of economic system should exist, questions on the environment and
    industrialism, and anarchists' roles in other movements.
  </p>
  The terms "anarchy" and "anarchism" are derived from the
  <collectionlink xlink:type="simple" xlink:href="11887.xml"> Greek </collectionlink>
- <emph2>
    <unknownlink src="Wiktionary:αναρχία"> αναρχία </unknownlink>
  </emph2>
  ("without
  <collectionlink xlink:type="simple" xlink:href="164556.xml"> archon </collectionlink>
  s (rulers)"). Thus "anarchism," in its most general meaning, is the belief that rulership is unnecessary and should be abolished. The word
  "anarchy", as most anarchists use it, does not imply
  <unknownlink src="chaos"> chaos </unknownlink>
```

The XML obtained

Fig. 1. Example of wiki → XML transformation for the *Anarchy* article (*12.xml*)

In order to uniformize this type of information, wikipedia uses templates. These templates are translated into XML using tags starting by *template_...* (for example : *template_country*). The template tags depend on the language of the collection because the templates are not the same depending on the language of the wikipedia collection used. An example of template is given in figure 2.

The Table 2 gives a summary of most frequent tags of the english collection.

DTD of the collection. Due to the irregularities of the structures of the documents, it is not possible to build a relevant DTD for the wikipedia XML corpus.

Table 2. Distribution and description of the main different XML node labels in the collection

Tag	Number of XML nodes	Description
collectionlink	≈ 17 millions	Hyperlink to a document of the collection
unknownlink	≈ 4 millions	Hyperlink to a document that does not exist in wikipedia
outsidelink	≈ 850,000	Hyperlink to a website
wikipedialink	≈ 850,000	Hyperlink to a wikipedia page (which is not in the collection)
languagelink	≈ 800,000	Hyperlink to the same page in an other laguage
emph2	≈ 2 millions	Emphasis level 2
emph3	≈ 1.5 millions	Emphasis level 3
emph4	≈ 1,000	Emphasis level 4
emph5	≈ 81,000	Emphasis level 5
table	≈ 92,000	Table
row	≈ 1 millions	Row of a table
cell	≈ 3.7 millions	Cell of a table
template	≈ 2.5 millions	Template tags
title	≈ 1.6 millions	Title of articles
p	≈ 2.8 millions	Paragraph
item	≈ 5.6 millions	Item of a list or enumeration
....	....	...

We give in Figure 3 an idea of the schema underlying the collection using a graphical representation of the tags inclusion.

General statistics about the Wikipedia XML collection are given in table 3

2.2 Categories

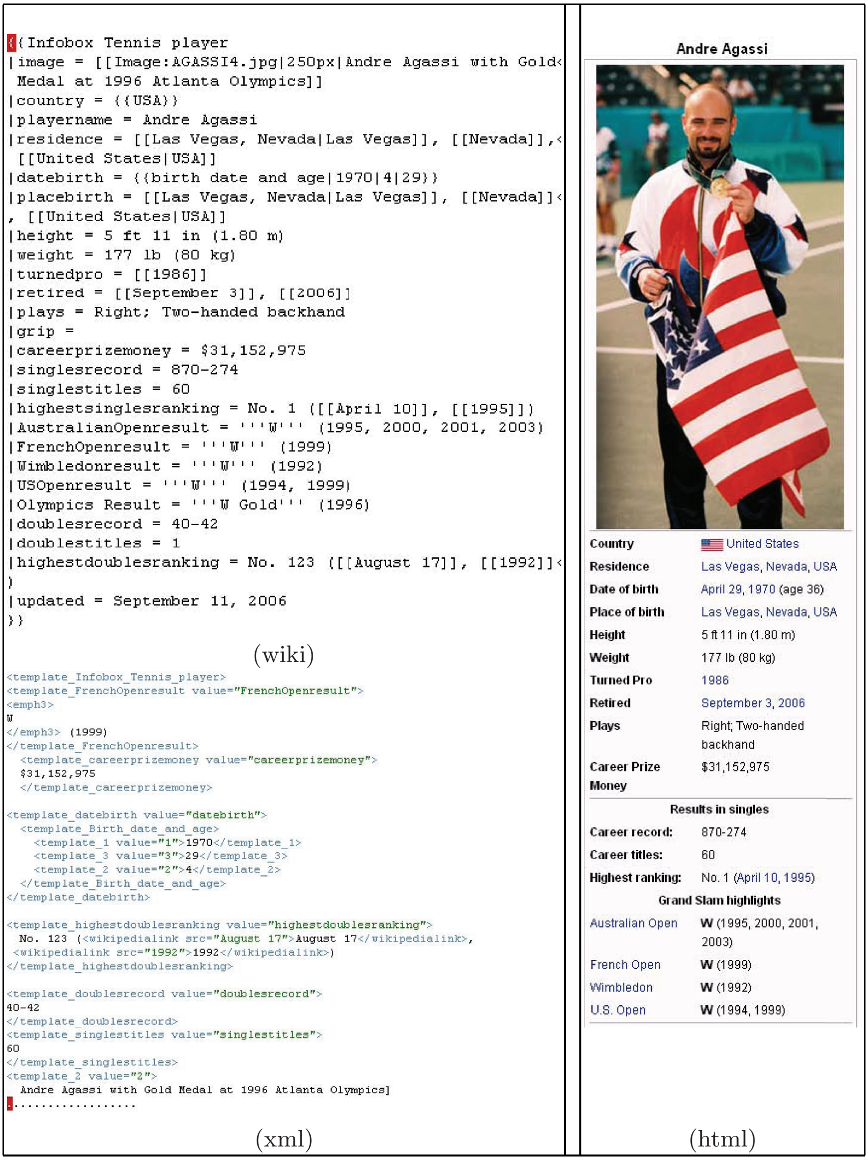
The documents of the wikipedia XML collections are organized in a hierarchy of categories defined by the authors of the articles. For each main collection, we propose a set of files describing:

- the hierarchy of categories (file : categories_hcategories.csv)
- the categories of each articles (file : categories_categories.csv)
- the categories names (file : categories_name.csv)

Table 3. Statistics about the structure of the documents from the *Main Collections*

Language	Mean size of document (bytes)	Mean Document Depth	Number of Nodes/Document
English	7,261	6.72	161.35
French	6,902	7.07	175.54
German	7,106	6.76	151.99
Dutch	5,092	6.41	122.8
Spanish	6,669	6.65	165.74
Chinese	6,664	6.91	179.23
Arabian	4,826	5.85	182.1
Japanese	7,973	7.1	94.96

Table 4 gives statistics about the categories.



```
{{Infobox Tennis player
|image = [[Image:AGASSI4.jpg|250px|Andre Agassi with Gold<
 Medal at 1996 Atlanta Olympics]]
|country = {{USA}}
|playername = Andre Agassi
|residence = [[Las Vegas, Nevada|Las Vegas]], [[Nevada]],<
 [[United States|USA]]
|datebirth = {{birth date and age|1970|4|29}}
|placebirth = [[Las Vegas, Nevada|Las Vegas]], [[Nevada]]<
, [[United States|USA]]
|height = 5 ft 11 in (1.80 m)
|weight = 177 lb (80 kg)
|turnedpro = [[1986]]
|retired = [[September 3]], [[2006]]
|plays = Right; Two-handed backhand
|grip =
|careerprizemoney = $31,152,975
|singlesrecord = 870-274
|singlestitles = 60
|highestsinglesranking = No. 1 ([[April 10]], [[1995]])
|AustralianOpenresult = '''W''' (1995, 2000, 2001, 2003)
|FrenchOpenresult = '''W''' (1999)
|Wimbledonresult = '''W''' (1992)
|USOpenresult = '''W''' (1994, 1999)
|Olympics Result = '''W Gold''' (1996)
|doublesrecord = 40-42
|doublestitles = 1
|highestdoublesranking = No. 123 ([[August 17]], [[1992]]<
)
|updated = September 11, 2006
}}
```

(wiki)

```
<template_Infobox_Tennis_player>
<template_FrenchOpenresult value="FrenchOpenresult">
<emph3>
W
</emph3> (1999)
</template_FrenchOpenresult>
  <template_careerprizemoney value="careerprizemoney">
  $31,152,975
  </template_careerprizemoney>

<template_datebirth value="datebirth">
  <template_Birth_date_and_age>
    <template_1 value="1">1970</template_1>
    <template_3 value="3">29</template_3>
    <template_2 value="2">4</template_2>
  </template_Birth_date_and_age>
</template_datebirth>

<template_highestdoublesranking value="highestdoublesranking">
  No. 123 (<wikipedialink src="August 17">August 17</wikipedialink>,
 <wikipedialink src="1992">1992</wikipedialink>)
</template_highestdoublesranking>

<template_doublesrecord value="doublesrecord">
40-42
</template_doublesrecord>
<template_singlestitles value="singlestitles">
60
</template_singlestitles>
<template_2 value="2">
  Andre Agassi with Gold Medal at 1996 Atlanta Olympics]
...................
```

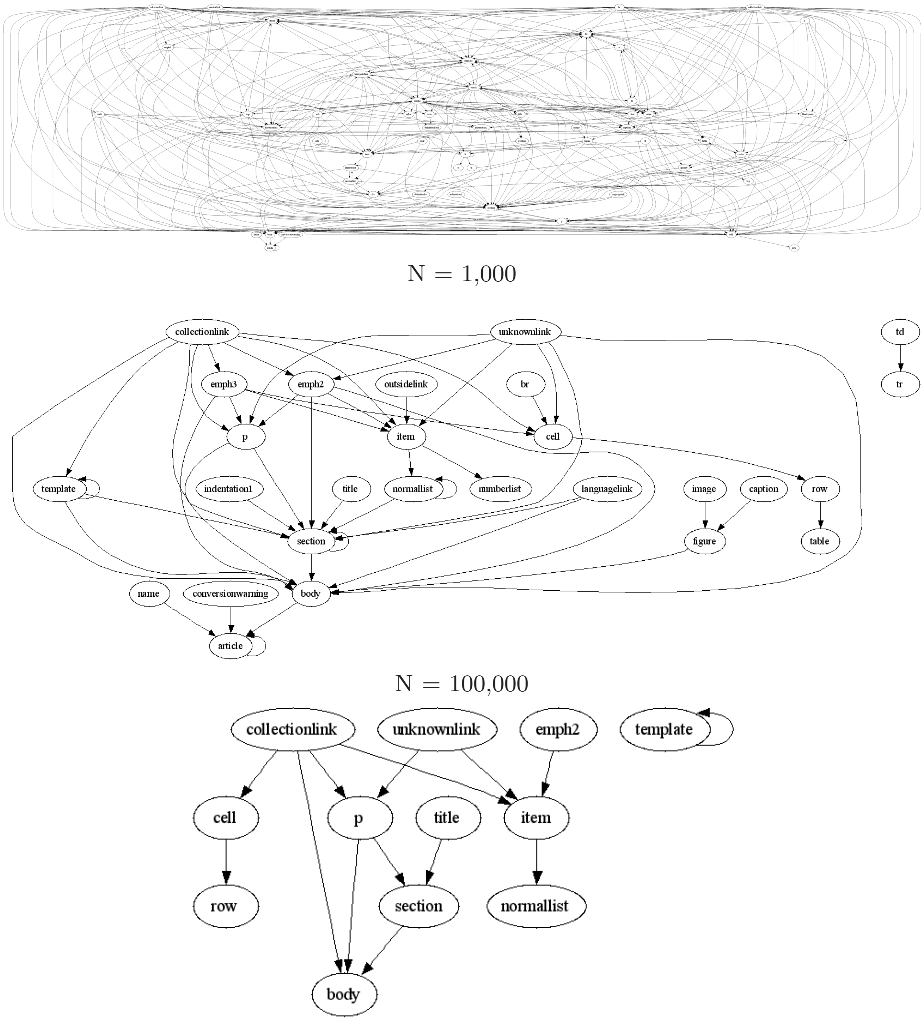
(xml)

Andre Agassi

Country	United States
Residence	Las Vegas, Nevada, USA
Date of birth	April 29, 1970 (age 36)
Place of birth	Las Vegas, Nevada, USA
Height	5 ft 11 in (1.80 m)
Weight	177 lb (80 kg)
Turned Pro	1986
Retired	September 3, 2006
Plays	Right; Two-handed backhand
Career Prize Money	$31,152,975
Results in singles	
Career record:	870-274
Career titles:	60
Highest ranking:	No. 1 (April 10, 1995)
Grand Slam highlights	
Australian Open	**W** (1995, 2000, 2001, 2003)
French Open	**W** (1999)
Wimbledon	**W** (1992)
U.S. Open	**W** (1994, 1999)

(html)

Fig. 2. Example of template conversion. A template (wiki) is converted using the *template_...* tags (xml). It correspond to a structured information of a wikipedia article (html).



N = 1,000

N = 100,000

N = 1,000,000

Fig. 3. This figure shows the schema of the documents (like a DTD). In this graph, each node corresponds to a possible XML node label. Two nodes (n_1, n_2) are connected if there exist at least N XML nodes with tag n_2 that have a child with tag n_1 in the corpus. For example, in the graph where N=1,000,000, the edge between *section* and *article* means that at least 1 million XML nodes with label *section* have a parent with label *article*.

Table 4. Statistics about the categories of the *Main Collections*

Language	Number of categories in the hierarchy	Mean number of categories for each document
English	113,483	2.2849
French	28,600	1.9570
German	27,981	2.5840
Dutch	13,847	1.6628
Spanish	12,462	1.6180
Chinese	27,147	2.0797
Japanese	26,730	2.0039

3 INEX 2006 Collections

3.1 Adhoc Collection

The collection used of the adhoc track during INEX 2006 is based on the english version of the wikipedia XML. Table 5 gives some general statistics over this collection and the Figure 4 gives the distribution of the documents with respect to their number of nodes.

Table 5. Statistics on the INEX 2006 AdHoc Corpus

Number of documents	659,388
Number of elements	≈ 52 millions
Size of the vocabulary	≈ 2 millions (depending on the preprocessing)
Number of tags	≈ 1,200

We do not detail here the description of the assesments made by the participants of INEX 2006.

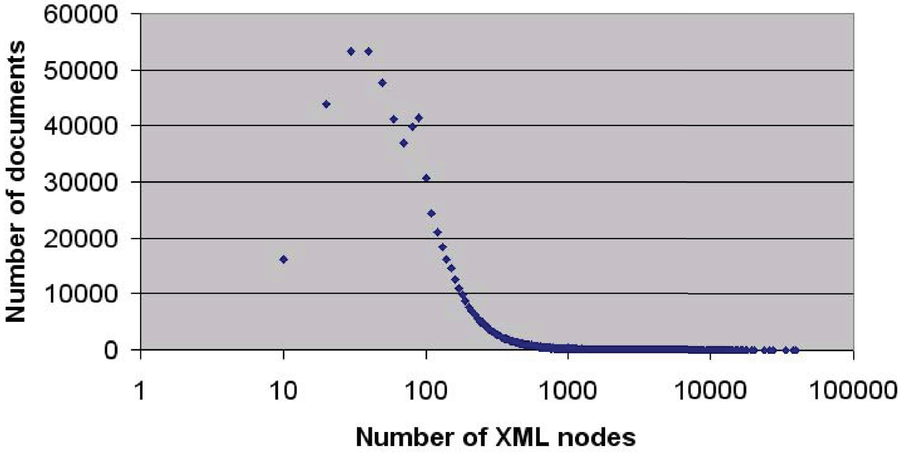


Fig. 4. Distribution of the documents wrt the number of doxels

Table 6. Statistics about the *XML Document Mining Challenge Collection* (*Single-Label Categorization Collection*)

Number of categories	60
Number of documents	150,094
Number of train documents	75,047
Number of test documents	75,047
Mean number of categories for each document	1
Structure of the corpus	The directory *documents* contains all the corresponding articles. The directory *relfiles* contains one file per category giving the id of the documents that belongs to this category[5].

3.2 XML Mining Track Collection

We provide a specific collection where each document belongs to **exactly** one category. This collection was used for the last yerar of the XML Mining Track. It is composed of the documents of the preceding collection belonging to a single category. This collection can be used for categorization and clustering of documents (see table 6). This collection is aimed at categorization/clustering benchmark.

4 Conclusion

This article report describes the main XML collections based on Wikipedia and developed for Structured Information Retrieval, Structured Machine Learning and Natural Language processing. Then, we detail the collection used during INEX 2006 for both the general adhoc retrieval track and the XML Mining track. Note that there exist some other corpus based on wikipedia XML for different tasks : Natural Language Processing, linguistic annotation, Question answering,etc.

Acknowlegment

The wikipediaXML corpus is distributed under the GPL Documentation license. It is completely free and can be used for non-profit educational and research purposes. All publications based on the wikipediaXML corpus should cite this article.