

VIETNAM NATIONAL UNIVERSITY – HO CHI MINH CITY
THE INTERNATIONAL UNIVERSITY
SCHOOL OF COMPUTER SCIENCE AND ENGINEERING



**IDENTIFICATION OF REPRESENTATIVE KEYWORDS
IN XML-BASED PATENTS**

By
Le Nguyen Thuy Hang

A thesis submitted to the School of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
Bachelor of Computer Science

Ho Chi Minh city, Vietnam
2017

IDENTIFICATION OF REPRESENTATIVE KEYWORDS IN XML-BASED PATENTS

APPROVED BY:

Nguyen Hong Quang, Ph.D.

THESIS COMMITTEE

ACKNOWLEDGMENTS

It is with deep gratitude and appreciation that I acknowledge the professional guidance of Dr. Nguyen Hong Quang. His constant encouragement and support helped me to achieve my goal.

TABLE OF CONTENTS

LIST OF TABLES	vi
LIST OF FIGURES	vii
ABSTRACT	viii
Chapter	
I. INTRODUCTION	1
II. LITERATURE REVIEW	3
1.Simple Statistics Approach	4
2.Linguistics Approach	5
3.Graph-based Approach.....	6
4.Hybrid Approaches.....	6
5.Summary	7
III. METHODOLOGY	8
1.Overview of Methodology	8
2.Detail of Methodology	9
2.1. Pre-processing	9
2.2. Semantic analysis	18
2.3. Term scoring and evaluation.....	19
IV. IMPLEMENTATION	22
1.Dataset Characteristics	22
2.Implementation Settings.....	25
3.Main Algorithms	26
3.1. N-gram phrase generation algorithm	26
3.2. Normalized Pointwise Mutual Information algorithm.....	27
3.3. TF-IDF weight algorithm.....	29
4.User Interface	31
V. RESULT / DISCUSSION	33
1.Result Description	33
2.Discussion	35
2.1. Result	35
2.2. Evaluation	37

VI. CONCLUSION	39
1.Summary	39
2.Future work	40
LIST OF REFERENCES	41
APPENDICES	43
Appendix A	43
Appendix B.....	48

LIST OF TABLES

Table 1. DOM methods example	10
Table 2. XPath expression used for Utility Patent	11
Table 3. Confusion Matrix for Keyword extraction.....	37

LIST OF FIGURES

Figure 1. Methodology Overview	8
Figure 2. DOM parser in JAVA	10
Figure 3. Setting threshold for term frequency in Phrase NPMI scoring.....	16
Figure 4. Sample WordNet synonym list	18
Figure 5. Tree view of XML-based Patent.....	23
Figure 6. Sample USPTO Patent from Google Patents search engine (Invention title and Abstract)	24
Figure 7. Sample USPTO Patent from Google Patents search engine (Description and Claims)	24
Figure 8. User Interface.....	31
Figure 9. Statistical Key term result	33
Figure 10. Key term result of patent in corpus containing patents of different classes	35
Figure 11. Key term result of patent in corpus containing patents of the same class	36
Figure 12. The first page of PDF Patent including Invention title and Abstract.....	43
Figure 13. The second page of PDF Patent	44
Figure 14. Figures included in PDF Patent	45
Figure 15. Description part of PDF Patent	46
Figure 16. Claims part of PDF Patent	47
Figure 17. Invention title part of XML-based Patent	48
Figure 18. Abstract part of XML-based Patent	49
Figure 19. Description part of XML-based Patent	49
Figure 20. Claims part of XML-based Patent	50

ABSTRACT

A patent is a detailed description of technology owned by its rightful owner and the information contained in it can bring high economic value. Identification of representative keywords of a patent can improve patent search and classification, leading to more successful protection and exploitation.

Current approaches of keyword identification simply extract keywords from text-based documents. Each document is structured in sections that carry different weights of importance (e.g., the Claim section is the core of a patent, hence it is the most important). However, using simple statistics approach or linguistics approach such as TF-IDF, WordNet, etc., separately has not fully exploited weight of each term in those sections and its linguistic features such as meaning or spelling, thus the keyword identification is not highly effective.

This research approach is to utilize XML parser, coordinate some Natural Language Processing (NLP) techniques logically to pre-process the content of the patent and handle its linguistics problem, before finally, applying weight to baseline TF-IDF model to fit content structure of the document. The final result of this approach is a set of meaningful and clean keywords which give high-level specification of the content of the document.

CHAPTER I

INTRODUCTION

A patent - a form of intellectual property - is an exclusive right granted by a country to an inventor, which forbids everyone except the inventor from making, using or selling his/her invention in that country during the time that patent is valid [20]. Filing patent applications early can limit the risk that someone else can get a patent on the same idea. If the inventor chooses not to exploit the invention, he/she can sell it or license the rights to commercialize it to another person or an organization, which can be a source of income.

One of the biggest sources providing various patents in the world is The United States Patent and Trademark Office (USPTO) [19] - the federal agency for granting U.S. patents and registering trademarks. Their patent is stored under XML-based format.

Despite of patents' importance, in reality, searching for the desired patents and classifying them is a hard job because of the large number of patent applications. Keywords play an important role in information retrieval; they can describe an entire document in just a few words. Therefore, it is necessary to develop an approach that can identify keywords automatically from the available data. Users can have a quick overview of those patents and improve the search engine of the patent storing system.

It is important to distinguish three terminologies used in this research: keyword, key phrase and key term. A keyword is a single word that describes or sums up the content of a document, e.g. "house", "shoes". In reality, keywords are actually

daily used terms which are familiar to every user. An upgraded version of keywords is called key phrases, which is a combination of single keywords, e.g., “white house”, “Nike shoes”. A key phrase can express the idea in a more detailed and precise way than a keyword. A term means a word or a group of words. Hence, key terms can be used as a replacement for either keywords or key phrases.

Therefore, this research aims to identify represent keywords from a patent document by exploiting XML-based content document, term-spelling ambiguity (e.g., the word “forget” can be written as “forget” , “forgot”, “forgotten” based on the tense of the sentence), diverse concept/meaning expression and unbalanced term weighting. Our contribution is threefold. First, our approach extracts key terms not from a normal text-based document but from a XML-based document by using the knowledge of parsing XML file. Second, utilize a few existed Natural Language Processing techniques such as WordNet, POS taggers, and coordinate them logically so that they can handle ambiguous term problems: term-spelling ambiguity and diverse meaning expression.

Most importantly, a patent is a document with an unbalanced section structure, which means different sections’ content carry different protection value. For example, the “Claims” part is the most important one because it is the legal grant of the patent to prevent others from stealing or using the invention, while the “Title” part is the most second important because it contains words which represents the main idea of the patent. Hence, dealing this problem by using and modifying the TF-IDF model with term weighting. This results in a user-friendly and fast application that can give out a set of meaningful keywords.

CHAPTER II

LITERATURE REVIEW

Keyword identification from a text data is a task using automatic extraction techniques to categorize and locate terms that best describe the main information of a document. [8, 9]

Key terms, key phrases or just keywords are the terminologies used for defining the terms that represent the topic discussed in a document. They not only provide the basic idea of the documents, but also help readers to search for further details more effectively or decide whether to continue further reading or not. Identifying keywords and key phrases from documents can help summarize contents explicitly, rapidly and concisely [10].

Keyword extraction task is a crucial problem in Text Mining, Information Retrieval and Natural Language Processing.

Methods for automatic keyword extraction can be categorized into:

- Supervised

Supervised methods require annotated data source [5]. A set of training data with labeled keywords is used to learn a model. The model is then applied to new set of documents to extract the keywords [21].

- Unsupervised

Unsupervised methods do not require annotations in advance [5].

Because Patent – XML based document is “unlabeled” data, the thesis focuses on the unsupervised keyword extraction methods.

Unsupervised methods are classified into below approaches:

1. Simple statistics approach
2. Linguistics-based approach
3. Graph-based approach
4. Hybrid approaches

1. **Simple Statistics Approach**

Methods of this approach are simple and do not need the training data. The statistical information of the words can identify the keywords in the document.

Commonly, keywords can be identified through the statistical information of the words in the text document or through non-linguistic features of the document that are concentrated such as the position of a word in the document, term frequency, etc.

Some of representative statistical methods include n-gram statistics, term frequency, word co-occurrence, TF-IDF model and PAT-tree (Patricia Tree; a suffix tree or a position tree). Their most common essential part depends on term frequency, it is the main criteria to decide whether a word is a keyword or not [4].

The method has a problem that it ignores words having low frequency even though they can be regarded as the keywords.

Using term frequency only may not bring out the most precise result as in a few text documents written in English or in any similar language in which the form of a word may vary. However, with a good pre-process stage, the evaluation based on term frequency is improved.

2. Linguistics Approach

These approaches pay their focuses on the linguistic features of sentences and documents to detect keyword in text documents. The linguistic approach includes the lexical analysis, syntactic analysis, semantic analysis, etc. Some resources used for lexical analysis are tree tagger, WordNet, N-gram, electronic dictionary. Similar methodologies/resources will be used for synaptic analysis: Noun phrase (NP), POS tagger, chunks (Parsing) [4].

Pudota et al. in [15] designs domain independent key phrase extraction system engaged N-grams, POS tags and statistics of each N-gram in defining candidate phrases [5].

Explanation [23]:

- Lexicon of a language is the words and phrases that it has. Lexical analysis is the act of identifying and analyzing the structure of a given text by diving it into smaller components such as paragraphs, sentences, words.
- Syntactic analysis, or parsing, means analyzing texts based on grammar and the relationship between words is formed by a certain way of arrangement. According to syntactic rules, sentences such as “the school goes to boy” is rejected.
- Semantic analysis means extracting the exact meaning/dictionary meaning of words from the text. The meaningfulness of the text is considered by drawing the syntactic structures and objects in the given text. Sentences such as “hot ice-cream” is wrong according to semantic rules.

3. *Graph-based Approach*

Graph-based text representation is considered as one of the best solutions to handle these problems [5, 14]. Graph is a mathematical model which allows users to explore the relationships and structural information.

Text document is modeled as graph where its vertices are terms and edges are relationship between edges. Below are some of principles exploited from different text scope and relations for graph construction, which can be used to establish the edges between two terms:

- Words co-occurrence in a sentence, paragraph, section or document added to the graph as a clique
 - Intersecting words from a sentence, paragraph, section or document
 - Words co-occurrence within the fixed window in text
 - Semantic relations: similar meaning, words spelled the same way but have different meaning, synonyms, antonyms, heteronyms, etc. [5]

TextRank is the most famous method for this unsupervised approach. Another related work that use this approach is KeyGraph uses community detection techniques for key terms extraction on Wikipedia's texts, modelled as a graph of semantic relationships between terms [5].

4. *Hybrid Approaches*

These approaches of keyword extraction mainly combine the methods mentioned above or use some heuristic knowledge in the task of keyword extraction on

the document. These sets of knowledge include position, length, layout feature of words, structure of documents, language, etc. [5].

5. Summary

The overview of the related works reveals that the thesis keyword extraction's approach is the hybrid approach. The approach includes identify and parse the right part in XML tag to get the wanted content, utilize the linguistic approaches such as POS tagger and WordNet dictionary to handle term-spelling ambiguity, diverse meaning expression and generate desired pattern N-gram phrase. Finally the term's position in the document will be identified and simple statistics approach TF-IDF algorithm for term scoring and evaluation will be improved.

CHAPTER III

METHODOLOGY

1. Overview of Methodology

Keyword identification is a big process containing many sub-processes to achieve the result – a key terms list that can describe a whole content of document in a few words. Below is the figure that describes step by step the full approach of keyword extraction.

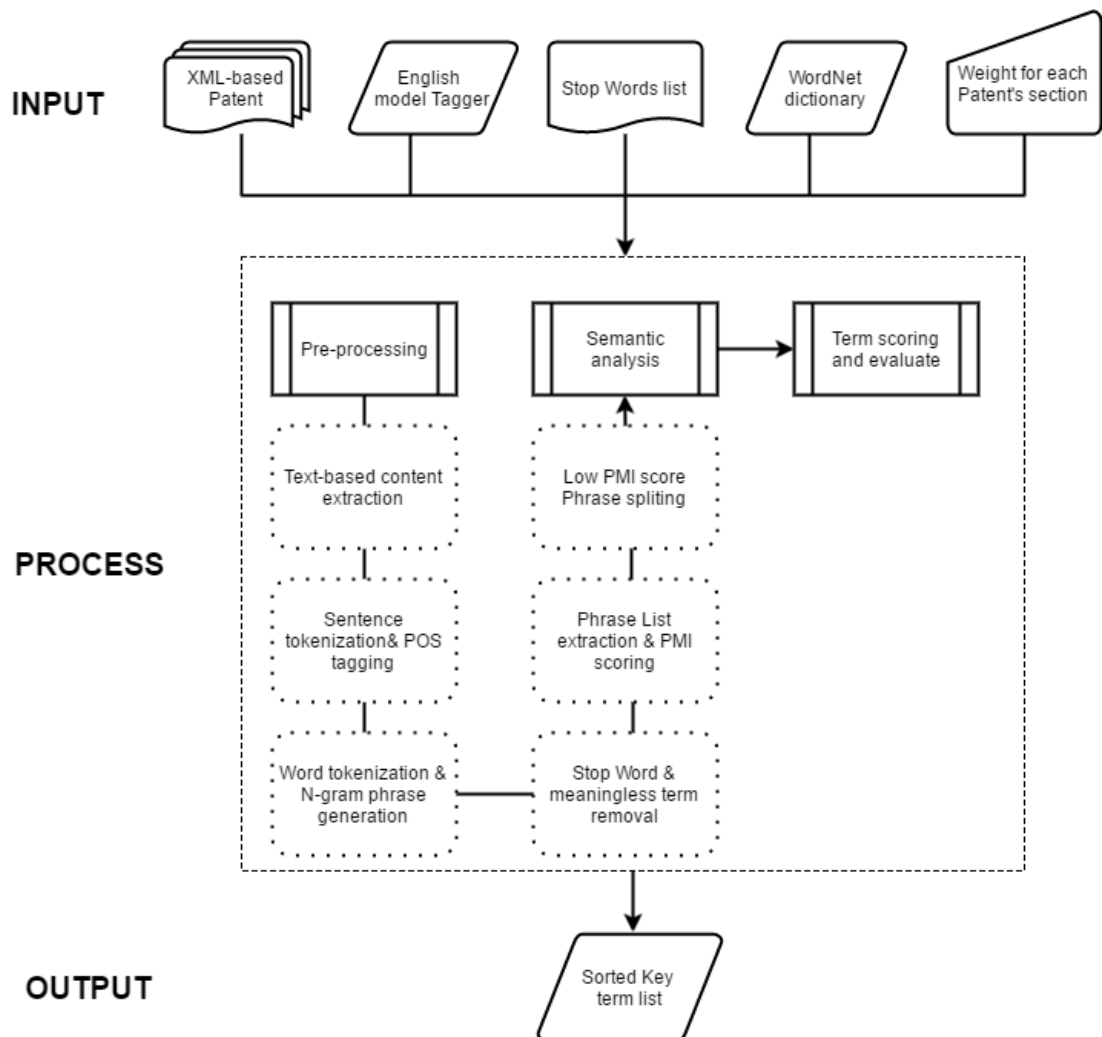


Figure 1. Methodology Overview

2. Detail of Methodology

This section gives a detailed description of all the steps mentioned in the methodology overview.

2.1.Pre-processing

Existing typical automatic keyword extractions only processes on the text document. Since the input patent document is in XML format and has words in bulk, it is necessary to pre-process the document.

2.1.1. Text-based content extraction

Patent documents is stored under XML format file. Before moving on to the next step, the most valuable data, which is text-based content, must be obtained. The solution for that is to use a parser to parse content in the XML tag.

DOM Parser

As the structure of the document is well known, DOM parser is a good parser to use. When parsing XML document by using DOM parser, a tree structure that contains all of the elements of the document is fully retrieved. DOM parser also provides some methods to examine the structure and content of the document.

DOM Methods	Result
Document.getDocumentElement()	Returns the root element of the document
Node.getFirstChild()	Returns the first child of a given Node
Node.getLastChild()	Returns the last child of a given Node.
Node.getNextSibling()	These methods return the next sibling of a given Node
Node.getPreviousSibling()	These methods return the previous sibling of a given Node
Node.getAttribute(attrName)	For a given Node, returns the attribute with the requested name

Table 1. DOM methods example

```

File file = new File("DTDS/" + name + ".xml");
DocumentBuilder dBuilder = DocumentBuilderFactory.newInstance()
    .newDocumentBuilder();
Document doc = dBuilder.parse(file);

```

Figure 2. DOM parser in JAVA

Nevertheless, Patent document is a XML complex tree (see Appendix B) contains many nodes while DOM is just a tree model of XML which provides low-level navigation capability (get First Child, get Sibling). Using DOM parser alone cannot fully retrieve the desired content.

Apply XPath to solve this problem since XPath has more high-level search navigation capability.

XPath

XPath is a major element in the XSLT standard. It is a query language that uses “path like” syntax to identify and navigate nodes in an XML document – XML

Path Language. XPath contains over 200 built-in functions so that users can utilize it to approach fully an XML file.

Since XPath uses path expressions to select nodes or node-sets in an XML document, these path expressions look the same as the expressions users see when working with a traditional computer file system. XPath expressions also can be used in many languages such as JavaScript, Java, XML Schema, PHP, Python, C and C++, etc.

Example of XPath applied for XML-based Patent:

XPath Expression	Result
<code>./invention-title</code>	Selects the invention title of the patent
<code>./abstract</code>	Selects the abstract section of the patent
<code>//heading[not(text()='CROSS-REFERENCE TO RELATED APPLICATIONS')]/following-sibling::p[contains(@id,'p-')]/node()[not(self::tables)]</code>	Select all sibling b of heading except heading content is 'CROSS-REFERENCE TO RELATED APPLICATIONS' and the children and grandchild – not tables node of that p.
<code>./tables/descendant-or-self::*text()</code>	Select text of descendant nodes and tables itself
<code>./claims</code>	Select text of descendant nodes and claims itself

Table 2. XPath expression used for Utility Patent

2.1.2. Phrase extraction and Evaluation

As mentioned above, the result is a set of key terms contains keywords and key phrases.

This session describes the approach of how to have meaningful phrases:

Instead of generating any N-grams phrase by just combining any two, three or more words which may not have any meaning, e.g. “of the machine”. The trend of phrases that can be potential key phrases is Noun or Adjective phrases. By applying Part-Of-Speech Tagger (POS Tagger)- a piece of software (JAVA implementation) that reads text in some language and assigns parts of speech to each word (and other token), such as noun, verb, adjective, etc., the task of identifying word class of each word becomes easier.

Below is the PENN Part of Speech Tags used in Stanford POS tagger:

1. CC	Coordinating conjunction	19. PRP\$	Possessive pronoun
2. CD	Cardinal number	20. RB	Adverb
3. DT	Determiner	21. RBR	Adverb, comparative
4. EX	Existential <i>there</i>	22. RBS	Adverb, superlative
5. FW	Foreign word	23. RP	Particle
6. IN	Preposition or subordinating conjunction	24. SYM	Symbol
7. JJ	Adjective	25. TO	<i>To</i>
8. JJR	Adjective, comparative	26. UH	Interjection
9. JJS	Adjective, superlative	27. VB	Verb, base form
10. LS	List item marker	28. VBD	Verb, past tense
11. MD	Modal	29. VBG	Verb, gerund or present participle
12. NN	Noun, singular or mass	30. VBN	Verb, past participle
13. NNS	Noun, plural	31. VBP	Verb, non-3rd person singular present
14. NNP	Proper noun, singular	32. VBZ	Verb, 3rd person singular present
15. NNPS	Proper noun, plural	33. WDT	Wh-determiner
16. PDT	Predeterminer	34. WP	Wh-pronoun
17. POS	Possessive ending	35. WP\$	Possessive wh-pronoun
18. PRP	Personal pronoun	36. WRB	Wh-adverb

Example:

- Original sentence: The present invention relates to conveyance rack for conveying metal rings
- POS tagged sentence: The_DT present_JJ invention_NN relates_VBZ to_TO conveyance_VB rack_NN for_IN conveying_VBG metal_NN rings_NNS

Using POS tagger helps extract N-gram phrase (a contiguous sequence of n single words) combining the following patterns:

- 2-gram phrase (a phrase contains 2 single words): Adjective + Noun, Noun + Noun, Adjective + Adjective, Proper Noun.
- 3-gram phrase (a phrase contains 3 single words): 2-gram phrase + Noun, 2-gram phrase + Adjective, Proper Noun.
- N-gram phrase (a phrase contains N single words): Proper Noun.

PMI (Pointwise Mutual Information)

After extracting the desired phrases, evaluating those phrases is necessary to keep the potential phrases and return the remaining into words if phrases' words components have more potential. Using Normalized Pointwise Mutual Information is a method to deal with this problem.

Definition

The Pointwise Mutual Information is usually used to compute the occurrence tendency of the two point wise events x and y together. Afterward, we can see if the result differs from what it should be when the two events are independent.

$$PMI(x, y) = \log_2 \frac{P(x, y)}{P(x)P(y)}$$

$PMI(x, y) = 0$: Particular values of x and y are statistically independent.

$PMI(x, y) > 0$: x and y co-occur more frequently than would be expected under an independence assumption

$PMI(x, y) < 0$: x and y co-occur less frequent than would be expected [12].

Application

In computational linguistics, PMI has been used for finding collocations and associations between terms of a text [11]. For instance,

If w_1 and w_2 represent for the first and second word respectively, instead of x and y , then the PMI for the two words w_1 and w_2 is given by:

$$PMI(w_1, w_2) = \log_2 \frac{P(w_1, w_2)}{P(w_1)P(w_2)}$$

Where $P(w_1, w_2)$ is the probability that two words w_1 and w_2 appears together in a certain text and $P(w_1)$ and $P(w_2)$ are the probabilities of w_1 and w_2 appearing separately in the text, respectively.

Moreover,

$$PMI(w_1, w_2) = \log_2 \frac{P(w_1, w_2)}{P(w_1) \times P(w_2)} = \log_2 \frac{\frac{C(w_1, w_2)}{\text{size}}}{\frac{C(w_1)}{\text{size}} \times \frac{C(w_2)}{\text{size}}}$$

The PMI score reaches its highest value when the number of occurrence of a phrase is equal with the number of occurrence of each single word builds up that phrase in the document.

$$P(w_1, w_2) = P(w_1) = P(w_2) \rightarrow C(w_1, w_2) = C(w_1) = C(w_2).$$

$$PMI(w_1, w_2) = \log_2 \frac{\text{size}}{C(w_1, w_2)}$$

As the formula shows, the rarer the phrase appears, the higher PMI score is. High PMI score does not mean high word dependence, it could also mean that the phrase is rarer.

An improved version of Pointwise Mutual Information is **Normalized Pointwise Mutual Information** which can reduce the impact of term frequency on ranking the phrase.

$$NPMI(w_1, w_2) = \frac{PMI(w_1, w_2)}{-\log_2 P(w_1, w_2)}$$

Result of **NPMI** is between [-1, +1] where -1 is for words never occurring together, 0 for independence, and +1 for complete co-occurrence [7].

Extending measure for tri-gram (3-gram) phrases [24]:

$$NPMI(w_1, w_2, w_3) = \frac{NPMI(w_1, w_2 w_3) + NPMI(w_1 w_2, w_3)}{2}$$

Threshold score of NPMI used for the demo is 0.7 (the number is gotten from tests and observation of statistical value of phrases)

Supposed that there are phrases and their word component appear only for a few times (one or two times) in the document, not to mention they always occur together. Their NPMI score is very high and term frequency is crucial in evaluating

their importance in the document. Therefore, high NPMI score does not indicate term's importance.

The NPMI score - 0.0 is set for phrases appear less than 5 times because these phrases can be useful in Semantic analysis and Term scoring after being split.

```

if (countS >= 5) {
    for (int i = 0; i < t.length; i++) {
        lower = lower * Proba(count.get(i), text.size());
    }

    upper = (double)Proba(countS, text.size() - 4);
    upper = (double)log2(upper / lower);
    result = (double)upper / (double) (-log2 (Proba(countS, text.size() - 4)));
    System.out.println(phrase + ":" + countS + ":" + count + ":" + text.size() + ":" + result);
}
if (countS < 5) {
    result = 0.0;
}
return result;

```

Figure 3. Setting threshold for term frequency in Phrase NPMI scoring

2.1.3. Phrase splitting

Before returning phrases into single words, because the desired result of previous step is achieved, the meaningless words or stop words (a set of commonly used words in any language, not just English. Because of their common existence in any document, they do not represent the subject of that document even though they occur a lot, e.g. “the”, “of”, “this”, etc.) must be eliminated.

The phrase splitting process seems to be easy because sketched task is to split the phrase in the content list if it matches with phrase in the Phrase split list. However, the comparison itself has unpredicted problems for following reasons:

- 3-gram phrase is a composite of two 2-gram phrase, e.g. “appropriate head gear” is a composite of “appropriate head” and “head gear”. The 3-gram phrase

has NPMI score below threshold but one of two 2-gram phrase NPMI is higher than NPMI threshold. As a result, separate one word from two other words make sense. E.g.

- Threshold score of NPMI: 0.7
- NPMI score of “appropriate head gear” : 0.5
- NPMI score of “appropriate head” : 0.8
- NPMI score of “head gear” : 0.4

Solution: get part of phrase needs to be split to compare. E.g., “head gear”.

- Low 3-gram NPMI score phrase also contains two 2-gram phrase whose NPMI are all higher than threshold. Hence, get the part of phrase only may split the same 2-gram stand-alone phrase. E.g.

- Threshold score of NPMI: 0.7
- NPMI score of “ocular gap size”: 0.5,
- NPMI score of “ocular gap”: 0.71
- NPMI score of “gap size”- 0.75.

Solution: Note part of the phrase needs to be split. E.g., “ocular gap size 1”.

- Low 2-gram NPMI score phrase exists in high 3-gram NPMI score phrase. E.g.

- Threshold score of NPMI : 0.7
- NPMI of “wireless HART network”: 0.8
- NPMI of “HART network” : 0.5

Solution: Note the 2-gram phrase. E.g., “HART network 0”.

The demo is implemented only to extract simple N-gram phrase like 2 or 3-gram phrase. In the future, if demo is implemented to extract longer phrase, there will be more complex problems.

2.2.Semantic analysis

In a document, many terms can express the same meaning or concept; people tend to use different words to avoid repetition in the content. E.g., “device” and “equipment”. The terms can be keywords but may be missed out because they appear less in the document.

WordNet is a large lexical database of English contains nouns, verbs, adjectives and adverbs. Those words are group into sets of cognitive synonyms (synsets) and each expresses a distinct concept. Synsets are interlinked based on of conceptual-semantic and lexical relations. WordNet is used to extract the synonym set of a term, which is then used to support its importance.

Taking advantages of POS tagger before, keeping the word class of each tagged word so that deriving term’s synonym set is more precise. This is because there are words which share the same spelling but belong to different classes. E.g., “shout”. Moreover, WordNet also helps to return the correct form of words. E.g., “describing” becomes “describe”.

```
run:
Verb@989103[describe,depict,draw] - give a description of
Verb@967067[report,describe,account] - to give an account or representation of in words
Verb@1585566[trace,draw,line,describe,delineate] - make a mark or lines on a surface
Verb@654017[identify,discover,key,key out,distinguish,describe,name] - identify as in botany or biology, for example
BUILD SUCCESSFUL (total time: 1 second)
```

Figure 4. Sample WordNet synonym list

2.3.Term scoring and evaluation

The subsection 2.3.1 below gives an overview of TF-IDF weight. In section 2.3.2, we describe how the TF-IDF weighting is modified to suit the content of the patent.

2.3.1. TF-IDF weighting

TF-IDF (term frequency – inverse document frequency) weight is a statistical measure, which is often used in information retrieval and text mining to evaluate the importance of a word/term in a document of a collection or a corpus. The importance of a word is directly proportional to the number of times it appears in the document and inversely proportional to the frequency of it in the corpus [13].

The TF-IDF weight is composed by two parts:

- Term Frequency (TF), as known as number of times a word appears in a document, divided by the total number of words in that document.

$$TF(t) = \frac{\text{Number of times term } t \text{ appears in a document}}{\text{Total number of terms in the document}}$$

- Inverse Document Frequency (IDF), computed as the logarithm of the number of the documents in the corpus divided by the number of documents where the specific term appears.

$$IDF(t) = \log \frac{\text{Total number of documents}}{\text{Number of documents with term } t \text{ in it}}$$

- $TF - IDF = TF \times IDF$
- TF-IDF weight is computed for each term in each document.

Example:

A document contains 100 words wherein the word “cat” appears 3 times.

The term frequency for term “cat”:

$$TF = 3 / 100 = 0.03$$

Assuming we have 10 million documents and the word “cat” appears in one thousand of these, the inverse document frequency is then calculated as:

$$IDF = \log (10,000,000 / 1,000) = 4$$

Thus, the TF-IDF weight is the product of these quantities:

$$TF-IDF = 0.03 * 4 = 0.12$$

2.3.2. Approach using the TF-IDF weight

After having a full pre-processed term list. The next step is to score each term to see how important they are. Why using TF-IDF? TF-IDF is a statistical measure, which can evaluate both aspects of a term. TF (term frequency) measure how frequently a term appears in a document while IDF measure its importance by measure its frequency outside the document. If a term appears less in other documents, it must represent the important information of evaluating document that other documents do not have.

Patent is an unbalanced structured section document. Therefore, location of a term in this kind document contributes to its importance. E.g., a term that appears in “Title” and “Claims” are more likely to become keyword. A solution for this is to add weight to term frequency. Weight varies from 0.0 (least important) to 1.0 (most important). The proposed TF formula for this type document as follows:

$$TF(t) = \frac{\Sigma(\text{Number of times } t \text{ appears in each section} \times \text{weight of each section})}{\Sigma(\text{Size of each section} \times \text{weight of each section})}$$

After term scoring by using modified TF-IDF algorithm, the final step is to evaluate which candidate terms can be the key terms of the patent document. Because the result of TF-IDF weight has already expressed the importance of each term, the evaluation's task is to rank the term based on its TF-IDF weight.

CHAPTER IV

IMPLEMENTATION

1. *Dataset Characteristics*

The input data set is downloaded from <https://bulkdata.uspto.gov/>

The data set tested in this thesis is **Utility Patent** (granted to anyone who invents or discovers any new and useful process, machine, article of manufacture, or composition of matter, or any new and useful improvement thereof.) [22] of the same day and picked randomly. Each document contains four main parts that we must focus on:

- Invention title
- Abstract
- Description
- Claims

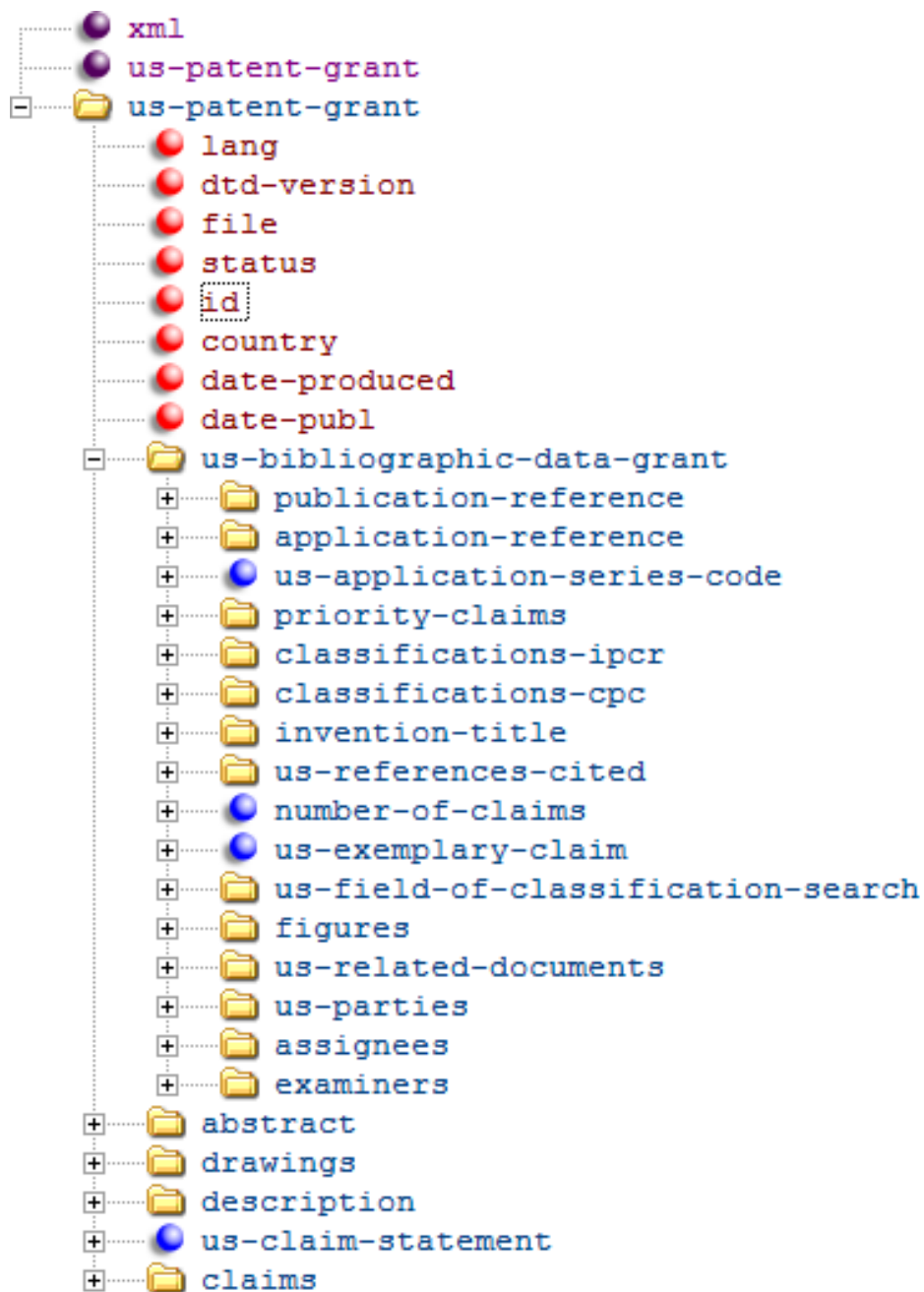


Figure 5. Tree view of XML-based Patent

Patents

Application
Grant

Invention title

Find prior art

Adjustable facial protector

US 8997266 B2

ABSTRACT

A head gear assembly that includes a shell; a facial protector connectively attached to the shell that further includes a first element; a second element positioned approximately parallel to the first element; a third and fourth element positioned approximately perpendicular to one or both of the first element and the second element; a gap further comprising a gap size defined by the position of a combination of at least two of the first element, the second element, the third element, and the fourth element, wherein the gap size is adjustable between a plurality of gap sizes, wherein the first element is movingly engaged with the second element, and wherein the first element moves freely from the second element as the gap size is adjusted.

Publication number	US8997266 B2
Publication type	Grant
Application number	US 14/187,524
Publication date	Apr 7, 2015
Filing date	Feb 24, 2014
Priority date	Dec 10, 2009
Also published as	US8695122, US20110138520, US20140165252, US20150173446
Inventors	John DeBoer
Original Assignee	John DeBoer
Export Citation	BIBTeX, EndNote, RefMan
Patent Citations	(47), Classifications (8)
External Links	USPTO, USPTO Assignment, Espacenet

Figure 6. Sample USPTO Patent from Google Patents search engine (Invention title and Abstract)

DESCRIPTION

CROSS-REFERENCE TO RELATED APPLICATIONS

This application is a continuation of U.S. Non-Provisional patent application Ser. No. 12/958,247, filed Dec. 1, 2010, which claims the benefit under 35 U.S.C. §119(e) of U.S. Provisional Patent Application Ser. No. 61/285,181, filed on Dec. 10, 2009. The disclosure of each application is hereby incorporated herein by reference in its entirety for all purposes.

STATEMENT REGARDING FEDERALLY SPONSORED RESEARCH OR DEVELOPMENT

Not applicable.

BACKGROUND OF THE DISCLOSURE

1. Field of the Disclosure

Embodiments disclosed herein relate generally to protective head gear. Other embodiments disclosed herein relate to protective headgear assembly for sports or activities generally associated with eye and/or facial protection as part of protective head gear. Specific embodiments disclosed herein may relate to protective sports equipment, and particularly to a facial protector used with a hockey helmet.

CLAIMS (16)

What is claimed:

1. A head gear assembly comprising:

- a shell configured for wearability on a users head;
- a facial protector connectively attached to the shell further comprising:
 - a first element;
 - a second element positioned approximately parallel to the first element;
 - a third and fourth element positioned approximately perpendicular to one or both of the first element and the second element;
 - a gap further comprising a gap size defined by the position of a combination of at least two of the first element, the second element, the third element, and the fourth element,

wherein the gap size is adjustable between a plurality of gap sizes, wherein the first element is movingly engaged with the second element, and wherein the first element moves freely from the second element as the gap size is adjusted.

Figure 7. Sample USPTO Patent from Google Patents search engine (Description and Claims)

2. Implementation Settings

The application is implemented in Java™ Platform (JDK 1.8). The demo contains below APIs and models.

- WordNet API (jaws-bin.jar)
- Stanford POS tagger API (stanford-postagger.jar)
- The English tagger model (english-bidirectional-distsim.tagger) and its properties file (english-bidirectional-distsim.tagger.props) for POS tagger API.
- English dictionary (WordNet-3.1 folder) for WordNet API.

3. Main Algorithms

Below is the pseudo code of three important algorithms used in the approach: N-gram phrase generation, Normalized Pointwise Mutual Information scoring used for evaluating phrases and TF-IDF weight for evaluating term's importance.

3.1. *N-gram phrase generation algorithm*

Algorithm 1: N-gram phrase generation
<p>Input: array String <i>sentence</i> Output : array String <i>term</i></p> <p>1 Initialize String <i>termT</i> = “ ”.</p> <p>2 For each word in the <i>sentence</i></p> <p> If word is Proper Noun</p> <p> Concatenate <i>termT</i> with word.</p> <p> If next word is not Proper noun, Noun or next word is Noun but length of <i>termT</i> ≥ 3.</p> <p> Add <i>termT</i> into <i>term</i>.</p> <p> Return <i>termT</i> empty.</p> <p> Go to next word.</p> <p> If word is Noun or Adjective</p> <p> Concatenate <i>termT</i> with word.</p> <p> If next word is not Noun, Adjective, Proper Noun, length of <i>termT</i> = 3 or next word is Proper Noun but Proper Noun's length ≥ 3 or 2 and <i>termT</i>'s length is already 1 or 2 respectively.</p> <p> Add <i>termT</i> into <i>term</i>.</p> <p> Return <i>termT</i> empty.</p> <p> Go to next word.</p> <p> Else</p> <p> <i>termT</i> = word.</p> <p> Add <i>termT</i> into <i>term</i>.</p> <p> Return <i>termT</i> empty.</p> <p> Go to next word.</p> <p>End for</p>

Input: variable *sentence* is a list of all single words of a sentence.

Output: variable *term* is a list of all term of a sentence including single words and phrases.

Block 2 of the algorithm describes how to generate the N-gram phrase following the desired pattern. The flow of the algorithm is to concatenate each word of the sentence with the next word if they match the pattern (Noun, Adjective or Proper Noun) and stop concatenating when phrases reach the desired length. Words which are Verb, Adverb or stand-alone Noun, Adjective remained as single word.

3.2. Normalized Pointwise Mutual Information algorithm

Algorithm 2: Normalized Pointwise Mutual Information
--

<p>Input: string <i>phrase</i>, array String <i>text</i></p>

<p>Output : double NPMI <i>result</i></p>
--

<p>1 Initialize array integer <i>count</i> of each <i>word</i> = 0 , integer <i>countS</i> of <i>phrase</i> =0</p>
--

<p>2 Split <i>phrase</i> into array <i>word</i>.</p>
--

<p>3 For each term in the <i>text</i></p>

<p> For each single word of <i>phrase</i></p>
--

<p> If term contains the word</p>
--

<p> Add one to <i>count</i> of the word.</p>

<p> If term contains <i>phrase</i></p>

<p> Add one to <i>countS</i> of the <i>phrase</i>.</p>

<p> End for</p>

<p>End for</p>

<p>4 If <i>countS</i> of the <i>phrase</i> ≥ 5 then</p>

<p> Calculate NPMI <i>result</i> based on <i>count</i> and <i>countS</i> of <i>phrase</i>.</p>

<p>Else</p>

<p> NPMI <i>result</i> the <i>phrase</i> = 0.</p>
--

Input: Variable *phrase* is a phrase in the Phrase list. Variable *text* is a list of all terms of the content.

Output: Variable *result* is the NPMI score of the input phrase.

Block 3 of the algorithm is a counter to count the appearance of each component word of phrase and the phrase itself in the patent document.

Block 4 is the NPMI score calculation of the phrase.

3.3.TF-IDF weight algorithm

Algorithm 3: TF-IDF

Input: array String *text*, array String 2d *term*, Object *Patent*

Output : Object *Patent*

- 1 Initialize integer array *count2* =0 for each *term*, integer array *count3* =0 for each *term*'s synonyms set, double array *result*, integer *weight*'s array *index*= 0, integer document *size* = 0, array String 2d *finalTerm* , array integer *docCount* = 0.
- 2 Add the main term into the each term list of 2d array *finalTerm*.
- 3 **For** each termText in the *text*.
 - If** the termText matches the separator for each part of the Patent.
 - Add one to the *weight array*'s *index* to get Patent's current part weight.
 - Else**
 - Add the weight of each part to the size of Patent
 - For** each term list in the 2d array *term*, each count2T and count3T of array *count2* and *count3*.
 - If** the main term in the term list matches the termText.
 - Add the current part weight to count2T.
 - If** the term's synonym part in the term list contains the termText.
 - Add the current part weight to count3T
 - Add the termText into the term list of the 2d array *finalTer*
 - End for**
- End for**
- 4 Delete *finalTerm* list elements if another *finalTerm* list elements contains them and reset the main term in the *finalTerm* list element to evaluate.
- 5 Add content of each available patent into array *docContent*.
- 6 **For** each content of *docContent*, each count of *docCount*.
 - For** each term list in 2d array *finalTerm*.
 - If** content contains the main term of each term list.
 - Add one to count.
 - End for**
- End for**
- 7 **For** each term list in the 2d array *finalTerm*, each countT, count2T and count3T of array *count*, *count2*, *count3* respectively.
 - countT = count2T + count3T.
 - Calculate TF-IDF score *result* using *size*, *docContent* size and each element of *count*, *docCount*.
- End for**
- 8 Sort the *finalTerm*, *result* in descending order using value of the *result*.
- 9 Set *result*, *finalTerm* for *Patent*.

Input: Variable *text* is a list of all terms of the content. Variable *term* is the list of sub-lists, each sub-list contains a candidate term and its synonym. Variable *Patent* is an object contains all the information of a patents : section weights, section content, patent ID.

Ouput: the object *Patent* with more information: sorted candidate terms, TF-IDF score of each candidate terms.

Block 2's function is to copy candidate terms from old list to final list. The old list and the final list is 2D lists which contain many sub-lists. Each sub-list consists of candidate term and its synonym.

Block 3 is the detailed counter to count the ocurrence of each candidate term and its synonyms in the document. The old list contains the candidate term and all its synonym extracted from WordNet dictionary. The candidate term's synonym is added into the final list along with the term if they appear in the content.

Block 4 of the algorithm is to arrange the final list again. The term-synonym list is deleted if other lists contain it. After deleting, the candidate term of each existing sub-list is also re-selected among these terms (the candidate term and its synonyms) based on the term frequency – the candidate term is the term with highest term frequency.

Block 5 is the function to get the content of other patents in corpus.

Block 6 is a counter to count the frequency of the term in the corpus.

Block 7 is the TF-IDF weight calculation for the sub-list 's candidate term of the final list. The calculation is based on the total number of occurrence of the term

and its synonym, the size of the patent content, the number of patents in the corpus and the frequency of the term in the corpus.

4. User Interface

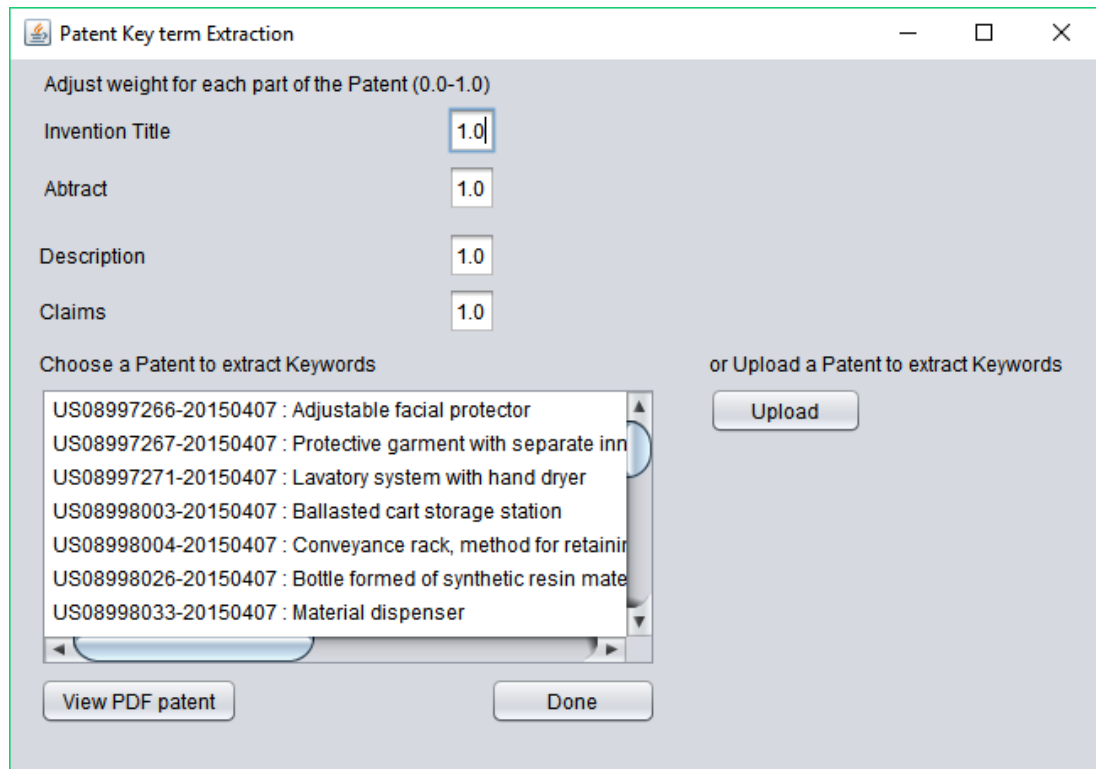


Figure 8. User Interface

The user interface contains:

- Available Patents in demo folder.
- View PDF Patent button: open the PDF Patent file for user to have a quick look at patent content (see Appendix A).
- Four input texts to adjust the weights for each part of the patent.
- Done button: start the automatic keyword extraction of the patent

- Upload button: upload a new XML-based patent to the system and start extract keywords from that patent

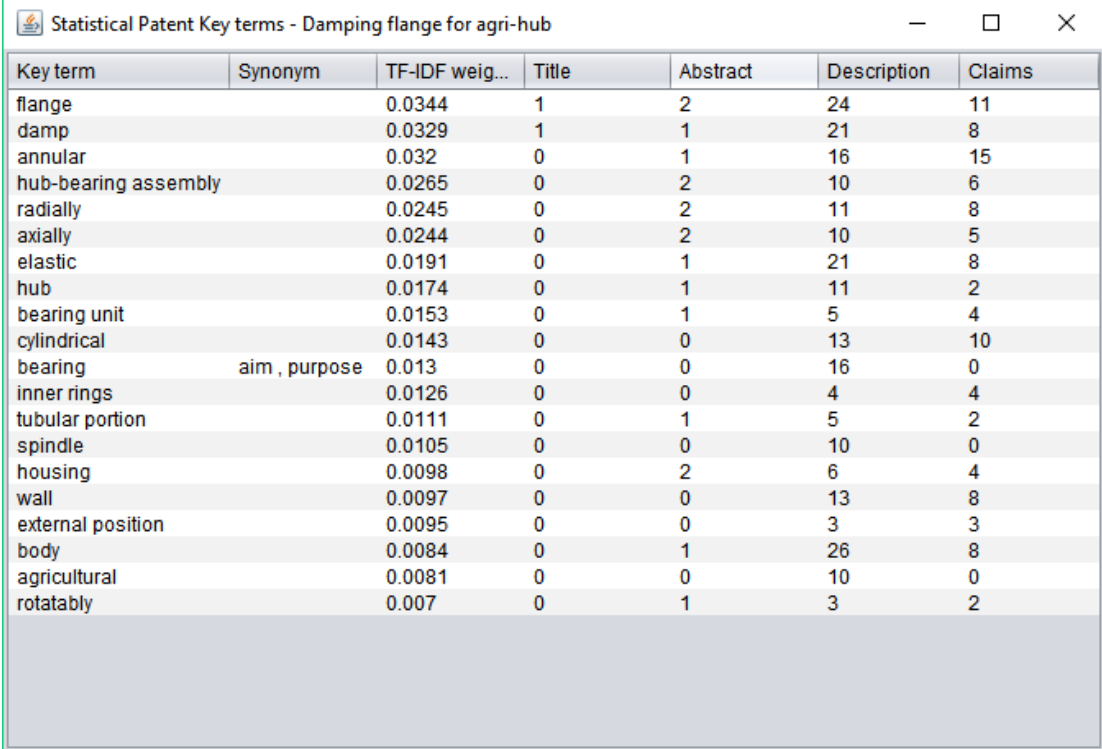
Input validation: the input texts to adjust the weights for each part of the patent must be in the range of 0.0 to 1.0

- 0.0 – least important
- 1.0 – most important

CHAPTER V

RESULT / DISCUSSION

1. Result Description



Key term	Synonym	TF-IDF weig...	Title	Abstract	Description	Claims
flange		0.0344	1	2	24	11
damp		0.0329	1	1	21	8
annular		0.032	0	1	16	15
hub-bearing assembly		0.0265	0	2	10	6
radially		0.0245	0	2	11	8
axially		0.0244	0	2	10	5
elastic		0.0191	0	1	21	8
hub		0.0174	0	1	11	2
bearing unit		0.0153	0	1	5	4
cylindrical		0.0143	0	0	13	10
bearing	aim , purpose	0.013	0	0	16	0
inner rings		0.0126	0	0	4	4
tubular portion		0.0111	0	1	5	2
spindle		0.0105	0	0	10	0
housing		0.0098	0	2	6	4
wall		0.0097	0	0	13	8
external position		0.0095	0	0	3	3
body		0.0084	0	1	26	8
agricultural		0.0081	0	0	10	0
rotatably		0.007	0	1	3	2

Figure 9. Statistical Key term result

By apply all sections with weight 1.0, user can have a key term extraction using baseline TF-IDF scoring.

The result includes four main points:

- The First column: main key term
- The Second column: term's synonyms in the document
- The Third column: TF-IDF score of each term

- Remaining columns: total number of terms and their synonyms occurring in each part – Invention Title, Abstract, Description, Claims.

The higher the TF-IDF score, the more significant word is. The keywords are sorted by TF-IDF score in descending order, so users can select as many keywords as they want from top to bottom. In the demo, we choose to display 20 highest ranking key terms. Calculating the total number of terms and its synonym that occurs in each part of the result is a way to ensure the pre-processing step does not miss any valuable term.

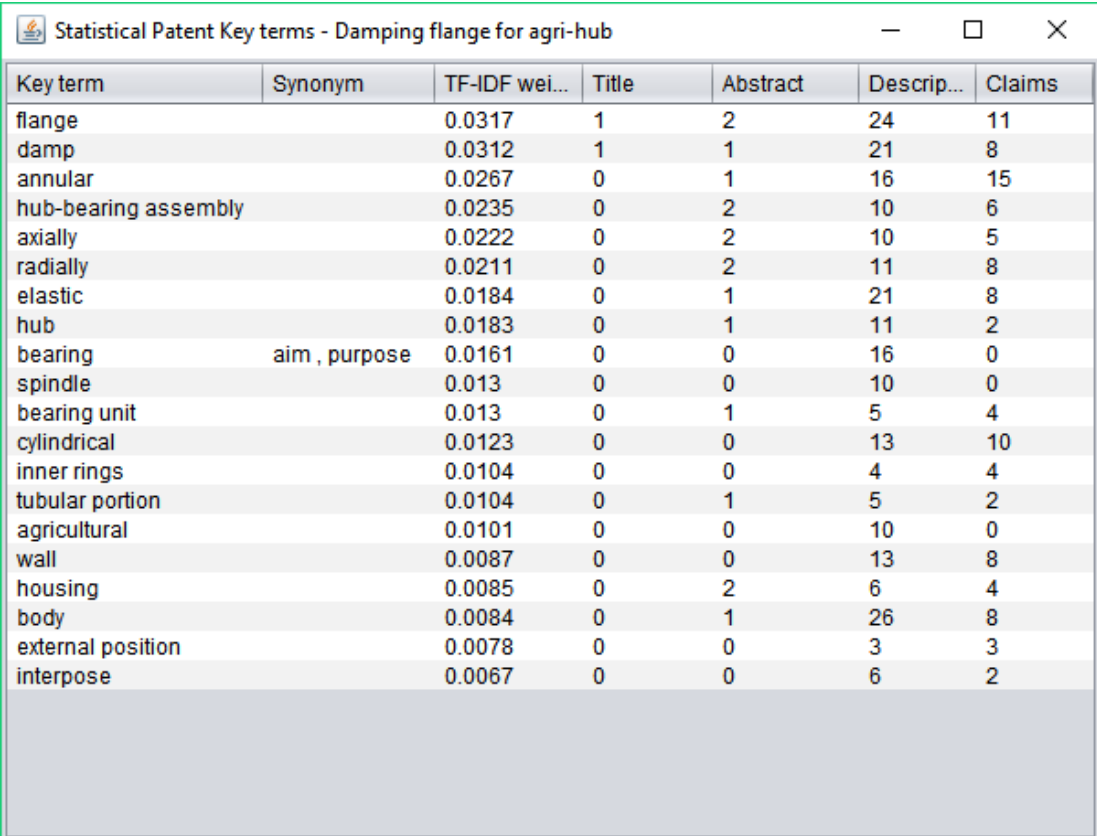
2. Discussion

2.1.Result

Below is the key term result of a patent processed in two different corpuses:

- The First corpus contains patents which belong to different classes.
- The Second corpus contains patents which belong to the same class. In

the sample result, the Class of the patents is “Agriculture”.



Key term	Synonym	TF-IDF wei...	Title	Abstract	Descrip...	Claims
flange		0.0317	1	2	24	11
damp		0.0312	1	1	21	8
annular		0.0267	0	1	16	15
hub-bearing assembly		0.0235	0	2	10	6
axially		0.0222	0	2	10	5
radially		0.0211	0	2	11	8
elastic		0.0184	0	1	21	8
hub		0.0183	0	1	11	2
bearing	aim , purpose	0.0161	0	0	16	0
spindle		0.013	0	0	10	0
bearing unit		0.013	0	1	5	4
cylindrical		0.0123	0	0	13	10
inner rings		0.0104	0	0	4	4
tubular portion		0.0104	0	1	5	2
agricultural		0.0101	0	0	10	0
wall		0.0087	0	0	13	8
housing		0.0085	0	2	6	4
body		0.0084	0	1	26	8
external position		0.0078	0	0	3	3
interpose		0.0067	0	0	6	2

Figure 10. Key term result of patent in corpus containing patents of different classes

Key term	Synonym	TF-IDF weig...	Title	Abstract	Description	Claims
damp		0.0276	1	1	21	8
flange		0.0272	1	2	24	11
annular		0.0229	0	1	16	15
elastic		0.0215	0	1	21	8
hub-bearing assembly		0.0214	0	2	10	6
cylindrical		0.0137	0	0	13	10
radially		0.0125	0	2	11	8
wall		0.0125	0	0	13	8
hub		0.0125	0	1	11	2
body		0.0124	0	1	26	8
axially		0.0122	0	2	10	5
bearing unit		0.0119	0	1	5	4
inner rings		0.0095	0	0	4	4
tubular portion		0.0095	0	1	5	2
spindle		0.0089	0	0	10	0
interpose		0.0071	0	0	6	2
external position		0.0071	0	0	3	3
bearing	aim , purpose	0.0067	0	0	16	0
housing		0.006	0	2	6	4
damages	damage , terms	0.0059	0	0	5	0

Figure 11. Key term result of patent in corpus containing patents of the same class

As mentioned before, TF-IDF weight evaluates the terms' importance based on its presence in a document (TF) and its rarity at a corpus level (IDF). Therefore, running the key term extraction for the same document in different corpus can lead to different results. Processing the sample patent in the corpus which contains other patents that have the same class "Agriculture" of it can lower the rank of some terms e.g., "agricultural". These terms may be the keywords of the Class.

Depend on the purpose of user, they can choose to run the application for the patent in different corpuses. Hence, if users want to identify keywords of Class of Patents, running each patent of the same Class in a corpus containing documents of different classes and finding the similar results between them can be a way to identify the keywords of the same class.

2.2. Evaluation

Keyword extraction is considered as a classification problem [17]. A document is a set of terms, which is classified into two categories: keyword and not keyword. Therefore, Confusion Matrix can be applied to summarize and evaluate the outcomes of automatic keyword extraction.

	Classified as a Keyword by Human annotator	Classified as NOT a Keyword by Human annotator
Classified as a Keyword by the approach	True Positive (TP)	False Positive (FP)
Classified as NOT a Keyword by the approach	False Negative (FN)	True Negative (TN)

Table 3. Confusion Matrix for Keyword extraction

Precision: exactness – what percentage of terms that the classifier labeled as key terms is actually key terms

$$Precision = \frac{TP}{TP + FP}$$

Recall: completeness – what percentage of key terms did the classifier label as key terms?

$$Recall = \frac{TP}{TP + FN}$$

F-measure: F measure (F1 or F-score): harmonic mean of precision and recall.

$$F = \frac{2 \times precision \times recall}{precision + recall}$$

Since there is no previous research to extract keyword of patent or human keyword extraction for this type of document- patent, there is no valuable resource to evaluate the result.

CHAPTER VI

CONCLUSION

1. Summary

Although keyword identification is a familiar process, applying it to XML-based Patent is quite new. Each type of document possesses different characteristics which means using the existed keyword extraction algorithms may not suitable for this type of document. Users can either keep the original TF-IDF model or adjust its section weights so that the algorithm is modified to suit the document's characteristics best.

Thanks to the usage of all existed NLP tools as well as how well the approach connects the tools are, the application is able to function smoothly, logically and give out a fine pre-processing and semantics analysis stage results which include the correct original spelling term, meaningful phrases and the related synonym sets. Nevertheless, it still suffers a few drawbacks. Since the tools do not possess fullest precision and the application is dependent on them to a point, the tools' minor mistakes or incompleteness may produce unwanted results.

Testing the application by comparing with the keywords assigned by users can help in examining how efficient the approach is. However, there is disadvantage. Patent is unsupervised data and there is none of related work before, so that there are no any resources of patent's keywords to fully evaluate the final result of the approach.

2. Future work

Even though there are remain several issues, this study can be the start to build a better and more completed automatic keyword identification for XML-based Patent.

Build benchmarking key term dataset by involving as many people as possible to read Patent and retrieving their suggestions in order to generate human-built key terms for the tools' evaluation.

The future application may obtain a few features including generating more complex N-gram phrases, extract phrase's synonyms in Semantic analysis step and using some Graph-based approach along with enhanced TF-IDF model to evaluate fully the importance of the word.

LIST OF REFERENCES

- [1] Bidyut Das, Subhajit Pal, Suman Kr. Mondal, Dipankar Dalui, Saikat Kumar Shome. *Automatic keyword extraction from any text document using N-gram Rigid Collocation*
- [2] P. Turney. *Extraction of Keyphrases from Text: Evaluation of Four Algorithms*
- [3] Santosh Kumar Bharti, Korra Sathya Babu, and Sanjay Kumar Jena. *Automatic Keyword Extraction for Text Summarization: A Survey*
- [4] Sifatullah Siddiqi, Aditi Sharan. *Keyword and Keyphrase Extraction Techniques: A Literature Review*
- [5] Slobodan Beliga. *Keyword extraction: a review of methods and approaches*
- [6] Fellbaum, Christiane (2005). *WordNet and wordnets*. In: Brown, Keith et al. (eds.), *Encyclopedia of Language and Linguistics, Second Edition*, Oxford: Elsevier, 665-670
- [7] Gerlof Bouma . *Normalized (Pointwise) Mutual Information in Collocation Extraction*
- [8] Beliga, Slobodan; Ana, Meštrović; Martinčić-Ipšić, Sanda. *An Overview of Graph-Based Keyword Extraction Methods and Approaches*.
- [9] Rada Mihalcea and Paul Tarau. *TextRank: Bringing Order into Texts*, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*
- [10] Beliga, Slobodan; Meštrović, Ana; Martinčić- Ipšić, Sanda. *Toward Selectivity-Based Keyword Extraction for Croatian News*
- [11] Paul L. Williams. *Information Dynamics: Its Theory and Application to Embodied Cognitive Systems*
- [12] Daniel Jurafsky, James H. Martin. *Speech and Language Processing. Chapter 18-Lexicons for Sentiment and Affect Extraction*
- [13] Anand Rajaraman, Jeffrey D. Ullman. *Mining of Massive Datasets*
- [14] S. S. Sonawane, Dr. P. A. Kulkarni. *Graph based Representation and Analysis of Text Document: A Survey of Techniques*

- [15] N. Pudota; A. Dattolo; A. Baruzzo; C. Tasso. “*A New Domain Independent Keyphrase Extraction System*” in *CCIS 2010, V.91*, pp. 67-78, 2010.
- [16] Alice Leung. *Evaluating Automatic Keyword Extraction for Internet Reviews*
- [17] Chengzhi ZHANG, Huilin WANG, Yao LIU, Dan WU, Yi LIA, Bo WANG. *Automatic Keyword Extraction from Documents Using Conditional Random Fields*
- [18] *Reasons for Patenting Your Inventions*. Wipo.int. Retrieved from http://www.wipo.int/sme/en/ip_business/importance/reasons.htm
- [19] *USPTO - United States Patents and Trademark Office*. Retrieved from <https://www.uspto.gov/>
- [20] *What is a patent?* Ipos.gov.sg. Retrieved from <https://www.ipos.gov.sg/AboutIP/TypesofIPWhatIsIntellectualProperty/WhatIsapatent.aspx>
- [21] Mohammad Rezaei, Najlah Gali, Pasi Fränti. *ClRank: A Method for Keyword Extraction from Web pages using clustering and distribution of nouns*
- [22] *General information concerning patents | USPTO*. Uspto.gov. Retrieved from <https://www.uspto.gov/patents-getting-started/general-information-concerning-patents>
- [23] *Artificial Intelligence Natural Language Processing*. www.tutorialspoint.com. Retrieved from https://www.tutorialspoint.com/artificial_intelligence/artificial_intelligence_natural_language_processing.htm
- [24] Sasa Petrovic, Jan Snajder, Bojana Dalbelo Basic, Mladen Kolar. *Comparison of Collocation Extraction Measures for Document Indexing*

APPENDICES

Appendix A

Sample PDF Patent


 US009538697B2	
<p>(12) United States Patent Ciulla et al.</p>	<p>(10) Patent No.: US 9,538,697 B2 (45) Date of Patent: Jan. 10, 2017</p>
<p>(54) DAMPING FLANGE FOR AGRI-HUB</p> <p>(71) Applicants: Luca Ciulla, Turin (IT); Carlo Maldera, Giaveno (IT)</p> <p>(72) Inventors: Luca Ciulla, Turin (IT); Carlo Maldera, Giaveno (IT)</p> <p>(73) Assignee: AKTIEBOLAGET SKF, Gothenburg (SE)</p> <p>(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.</p> <p>(21) Appl. No.: 14/715,644</p> <p>(22) Filed: May 19, 2015</p> <p>(65) Prior Publication Data US 2015/0327427 A1 Nov. 19, 2015</p> <p>(30) Foreign Application Priority Data May 19, 2014 (IT) TO2014A0393</p> <p>(51) Int. Cl. <i>F16C 35/07</i> (2006.01) <i>A01B 71/04</i> (2006.01) <i>F16C 19/54</i> (2006.01) <i>F16C 33/60</i> (2006.01) <i>F16C 19/18</i> (2006.01)</p> <p>(52) U.S. Cl. CPC <i>A01B 71/04</i> (2013.01); <i>F16C 35/07</i> (2013.01); <i>F16C 19/184</i> (2013.01); <i>F16C 2226/62</i> (2013.01); <i>F16C 2310/00</i> (2013.01)</p> <p>(58) Field of Classification Search CPC F16C 19/08; F16C 19/184; F16C 35/07; F16C 2310/00; F16C 2226/62; F16C 2208/18; A01B 15/16; A01B 71/04; B60B 27/0005</p>	<p>USPC 384/296-297, 460, 492, 504, 536, 547, 384/544, 589; 172/394, 604 See application file for complete search history.</p> <p>(56) References Cited U.S. PATENT DOCUMENTS</p> <p>1,584,616 A * 5/1926 Cothran A01B 23/06 384/460 1,701,518 A * 2/1929 Walker F16F 15/1442 74/574.4 2,299,010 A * 10/1942 Doman F01P 5/02 416/134 R 2,698,565 A * 1/1955 Carney A01B 39/14 172/574 2,961,894 A * 11/1960 Oles F16F 15/1442 29/450 3,861,828 A * 1/1975 Biermann B64C 11/008 416/145 4,252,385 A * 2/1981 Leitzel F16C 33/74 384/138</p> <p style="text-align: center;">(Continued)</p> <p style="text-align: center;">FOREIGN PATENT DOCUMENTS</p> <p>BE 368492 A 5/1936 DE 20012666 U1 12/2000</p> <p style="text-align: center;">(Continued)</p> <p><i>Primary Examiner</i> — Marcus Charles (74) <i>Attorney, Agent, or Firm</i> — Bryan Peckjian; SKF USA Inc. Patent Dept.</p> <p>(57) ABSTRACT A hub-bearing assembly for rotatably mounting a tilling disc about an axis of rotation. The hub-bearing assembly includes an annular hub providing an axially extending tubular portion, comprising a housing and a radially outer flange for mounting a disc. A bearing unit is mounted within the housing. An elastic damping body axially is fitted between the radially outer flange and the disc.</p> <p style="text-align: center;">6 Claims, 3 Drawing Sheets</p>

Figure 12. The first page of PDF Patent including Invention title and Abstract

(56)

References Cited

U.S. PATENT DOCUMENTS

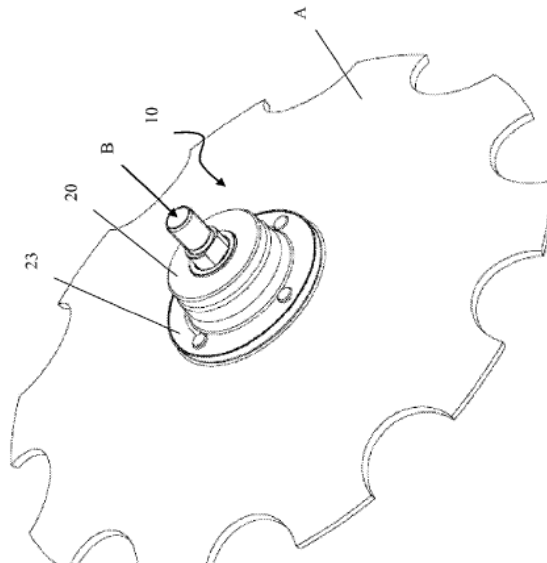
4,295,691 A * 10/1981 Rubenthaler F16F 1/3732
 384/297
 5,802,995 A * 9/1998 Baugher A01C 7/205
 111/140
 6,082,276 A * 7/2000 Klein A01C 5/064
 111/164
 6,364,426 B1 * 4/2002 Horne B24B 5/065
 384/544
 7,073,949 B2 * 7/2006 Ruckle A01B 71/04
 384/460
 7,475,738 B2 * 1/2009 Frasier A01B 71/04
 111/140
 8,397,602 B2 * 3/2013 Christenson F16F 15/126
 74/574.4
 8,899,345 B2 * 12/2014 Lazcano Lasa A01B 23/06
 172/604
 2003/0201108 A1 * 10/2003 Linden F16C 35/06
 172/604
 2007/0147719 A1 * 6/2007 Komori B60B 27/00
 384/492
 2012/0045155 A1 * 2/2012 Morero A01B 71/04
 384/480
 2014/0029885 A1 * 1/2014 Ciulla A01B 71/04
 384/536

FOREIGN PATENT DOCUMENTS

EP 2689648 A1 1/2014
 JP 02095152 A * 4/1990
 JP 07026950 A * 1/1995
 JP 2000074127 A1 * 3/2000 B60K 5/12
 WO WO 0219791 A1 * 3/2002 A01B 15/16
 WO WO 02070285 A1 * 9/2002 B25B 27/0035
 WO 2007105185 A2 9/2007

* cited by examiner

Figure 13. The second page of PDF Patent

**Fig. 2*****Figure 14. Figures included in PDF Patent***

1

DAMPING FLANGE FOR AGRI-HUB

CROSS REFERENCE TO RELATED APPLICATIONS

This is a Non-Provisional Patent Application, filed under the Paris Convention, claims the benefit of Italy Patent (IT) Application Number TO2014A000393 filed on 19 May 2014, which is incorporated herein by reference in its entirety.

TECHNICAL FIELD

The present invention is related to a hub-bearing assembly for an agricultural tilling disc.

As known, discs for agricultural use are usually rotatably assembled on corresponding spindle, projecting from the frame of a plough or other agricultural machine.

BACKGROUND ART

From document WO 2002/019791 A is known a hub-bearing assembly for rotatably mounting a tilling disc for agricultural use, around an axis of rotation. The assembly comprises an annular hub, having a tubular portion axially extending, which defines a substantially cylindrical housing and a radially outer flange for fixing the disc. In the housing a bearing unit is located, the bearing unit comprising an outer ring, one or two inner rings and one or two set of rolling bodies, interposed between inner and outer rings. In other solutions, the outer ring is in one piece with the flanged hub.

During the working life, impacts of the disc against stones and similar bodies damage the bearing raceways, reducing the bearing lifetime.

BRIEF SUMMARY OF THE INVENTION

2

FIG. 3 is an enlarged view, partly in cross section, of the hub-bearing assembly and the spindle of FIG. 1; and FIG. 4 is an enlarged detail of the cross section in FIG. 3.

DETAILED DESCRIPTION OF THE INVENTION

With reference to the above figures, a hub-bearing assembly according to an embodiment of the invention, referenced as a whole with **10**, is used for mounting a tilling disc **A** in a freely rotatable way around an axis of rotation **x**, which is defined by a spindle **B** projecting from an agricultural machine or tool (not shown), as an example, a plow, a harrow or other similar tools. Features of disc **A**, which can be a whatever known disc, for example a disc for plowing or a disc for seeding (suitable for opening furrows in a previously plowed land), are not relevant for the invention understanding and therefore will not be described in further details.

With reference to FIG. 3, the hub-bearing assembly **10** comprises a hub **20**, a bearing unit **30** housed in the hub **20** and an elastic damping body **40**, axially interposed between the hub and the disc **A**.

In particular, the hub **20** has a substantially annular shape and presents a main tubular portion **21** axially extended, which internally defines a substantially cylindrical housing **22** for the bearing unit **30**. The housing is radially confined by an inner wall **22a** substantially cylindrical. Throughout the present description and in the claims, the terms and expressions indicating positions and orientations such as "radial" and "axial" are to be taken to refer to the axis of rotation **x** of the bearing unit **30**.

From a first axial end of the tubular portion **21** of the hub, a radially outer flange **23** extends, the flange having a plurality of axial holes for mounting the disc **A** by means of suitable fastening means, for example screws **24**. From a second axial end of the tubular portion **21** a radially inner flange **25** extends, axially confining the housing **22** on the

Figure 15. Description part of PDF Patent

surface Aa of the disc A.

The cylindrical wall 42 of the elastic damping body 40 is radially coupled with the flange 23, by means of a substantially cylindrical surface 42a of the cylindrical wall 42, in a radially internal position, and a corresponding substantially cylindrical surface 23b of the flange 23, in a radially external position.

The elastic damping body 40 is coupled to the flange 23 by means of known processes. As an example, the coupling between the two components can be realized co-molding the elastic damping body 40 on the flange 23 or gluing them or connecting them by mechanical fastening means.

Preferably, the substantially cylindrical surface 23b of the flange 23 is stepwise shaped, in other words is formed by a plurality of cylindrical portions 23b', 23b'', 23b''', having decreasing diameters and each other connected by annular surfaces 23c', 23c''. As a consequence, also the surface 42a of the cylindrical wall 42 of the elastic damping body 40 is stepwise shaped as well. In this way, the coupling surface between elastic damping body and flange increases; moreover, the stepwise shape of the two coupling surfaces improves the grip between elastic body and flange and consequently a more stable coupling between them is obtained.

Advantageously, the flange 23 is provided with an annular groove 23d, located along the substantially annular surface 23a and consequently the annular wall 41 of the elastic damping body 40 has an annular protrusion 41a, which is coupled with the annular groove 23d. In this way, the thickness of the elastic damping body can be increased, obtaining at the same time a greater damping effect and a better grip in the coupling elastic body-flange.

In definitive, the elastic body 40 interposed between the disc and the flange can absorb vibrations and hits, which derive by the impact of the disc against stone materials on the agricultural land. In practice, the elastic damping body works as a shock absorber. In fact all impulsive load due to

various changes may be made in the function and arrangement of elements described in an exemplary embodiment without departing from the scope as set forth in the appended claims and their legal equivalents.

The invention claimed is:

1. A hub-bearing assembly for rotatably mounting a tilling disc about an axis of rotation, the assembly comprising:
 - an annular hub providing an axially extending tubular portion, comprising a housing and a radially outer flange for mounting a disc;
 - a bearing unit mounted within the housing, the bearing unit comprising:
 - an outer ring,
 - a pair of inner rings, and
 - a dual set of rolling elements, interposed between the outer ring and the pair of inner rings; and
 - an elastic damping body axially fitted on the radially outer flange, on the side facing the mountable tilling disc, the elastic damping body further comprising a substantially annular wall, the substantially annular wall is steadily connected to a substantially cylindrical wall, the cylindrical wall being in a radially external position respect to the annular wall, the elastic damping body is fabricated of one of an elastomeric material or a plastic material, wherein the cylindrical wall of the elastic damping body is radially coupled with the flange, by:
 - a substantially cylindrical surface of the cylindrical wall, in a radially internal position, and
 - a corresponding substantially cylindrical surface of the flange, in a radially external position.
2. A hub-bearing assembly according to claim 1, wherein the elastic damping body is located on a substantially annular surface of the flange, in an axially external position.
3. A hub-bearing assembly according to claim 2, the flange further comprising an annular groove, wherein the annular groove is located along the substantially annular

Figure 16. Claims part of PDF Patent

Appendix B

Sample XML-based Patent

```
<main-group>2310</main-group>
<subgroup>00</subgroup>
<symbol-position>L</symbol-position>
<classification-value>A</classification-value>
<action-date><date>20170110</date></action-date>
<generating-office><country>US</country></generating-office>
<classification-status>B</classification-status>
<classification-data-source>H</classification-data-source>
<scheme-origination-code>C</scheme-origination-code>
</classification-cpc>
</further-cpc>
</classifications-cpc>
<invention-title id="d2e61">Damping flange for agri-hub</invention-title>
<us-references-cited>
<us-citation>
<patcit num="00001">
<document-id>
<country>US</country>
<doc-number>1584616</doc-number>
<kind>A</kind>
<name>Cothran</name>
<date>19260500</date>
</document-id>
</patcit>
<category>cited by examiner</category>
<classification-cpc-text>A01B 23/06</classification-cpc-text>
<classification-national><country>US</country><main-classification>384460<
</us-citation>
<us-citation>
```

Figure 17. Invention title part of XML-based Patent


```

</assignee>
</assignees>
<examiners>
  <primary-examiner>
    <last-name>Charles</last-name>
    <first-name>Marcus</first-name>
    <department>3656</department>
  </primary-examiner>
</examiners>
</us-bibliographic-data-grant>
<abstract id="abstract">
  <p id="p-0001" num="0000">A hub-bearing assembly for rotatably mounting a tilling disc about an ax
includes an annular hub providing an axially extending tubular portion, comprising a housing and a
disc. A bearing unit is mounted within the housing. An elastic damping body axially is fitted betw
disc.</p>
</abstract>
<drawings id="DRAWINGS">
  <figure id="Fig-EMI-D00000" num="00000">
    <img id="EMI-D00000" he="196.34mm" wi="330.03mm" file="US09538697-20170110-D00000.TIF" alt="embedd
="tif"/>
  </figure>
  <figure id="Fig-EMI-D00001" num="00001">
    <img id="EMI-D00001" he="249.94mm" wi="196.09mm" orientation="landscape" file="US09538697-20170110
img-content="drawing" img-format="tif"/>
  </figure>
</drawings>

```

Figure 18. Abstract part of XML-based Patent

```

<?BRFSUM description="Brief Summary" end="lead"?>
<heading id="h-0001" level="1">CROSS REFERENCE TO RELATED APPLICATIONS</heading>
<p id="p-0002" num="0001">This is a Non-Provisional Patent Application, filed under the Paris
Patent (IT) Application Number TO2014A000393 filed on 19 May 2014, which is incorporated herei
<heading id="h-0002" level="1">TECHNICAL FIELD</heading>
<p id="p-0003" num="0002">The present invention is related to a hub-bearing assembly for an ag
<p id="p-0004" num="0003">As known, discs for agricultural use are usually rotatably assembled
the frame of a plough or other agricultural machine.</p>
<heading id="h-0003" level="1">BACKGROUND ART</heading>
<p id="p-0005" num="0004">From document WO 2002/019791 A is known a hub-bearing assembly for r
agricultural use, around an axis of rotation. The assembly comprises an annular hub, having a
defines a substantially cylindrical housing and a radially outer flange for fixing the disc. I
bearing unit comprising an outer ring, one or two inner rings and one or two set of rolling bo
rings. In other solutions, the outer ring is in one piece with the flanged hub.</p>
<p id="p-0006" num="0005">During the working life, impacts of the disc against stones and simi
reducing the bearing lifetime.</p>
<heading id="h-0004" level="1">BRIEF SUMMARY OF THE INVENTION</heading>
<p id="p-0007" num="0006">Aim of the present invention is to realize a hub-bearing assembly fo
overcomes the above mentioned inconveniences.</p>
<p id="p-0008" num="0007">This and other purposes and advantages, which will be better hereaft
aspect of the invention by a hub-bearing assembly as defined in the enclosed independent claim
<p id="p-0009" num="0008">Further embodiments of the invention, preferred and/or particularly
characteristics as in the enclosed dependent claims.</p>
<p id="p-0010" num="0009">In practice, the hub-bearing assembly comprises an elastic damping b
disc for agricultural use. The elastic damping body absorbs part of dynamic loads, due to the
Therefore, the dynamic loads no more fully transmitted to the bearing and its rolling bodies,
</p>
<?BRFSUM description="Brief Summary" end="tail"?>
<?brief-description-of-drawings description="Brief Description of Drawings" end="lead"?>
<description-of-drawings>

```

Figure 19. Description part of XML-based Patent

```

<?DETDESC description="Detailed Description" end="tail"?>
</description>
<us-claim-statement>The invention claimed is:</us-claim-statement>
<claims id="claims">
<claim id="CLM-00001" num="00001">
<claim-text>1. A hub-bearing assembly for rotatably mounting a tilling disc about an axis of
<claim-text>an annular hub providing an axially extending tubular portion, comprising a hous
disc;</claim-text>
<claim-text>a bearing unit mounted within the housing, the bearing unit comprising:
<claim-text>an outer ring,</claim-text>
<claim-text>a pair of inner rings, and</claim-text>
<claim-text>a dual set of rolling elements, interposed between the outer ring and the pair o
</claim-text>
<claim-text>an elastic damping body axially fitted on the radially outer flange, on the side
damping body further comprising a substantially annular wall, the substantially annular wall
cylindrical wall, the cylindrical wall being in a radially external position respect to the
fabricated of one of an elastomeric material or a plastic material, wherein the cylindrical
coupled with the flange, by:
<claim-text>a substantially cylindrical surface of the cylindrical wall, in a radially inter
<claim-text>a corresponding substantially cylindrical surface of the flange, in a radially e
</claim-text>
</claim-text>
</claim>
<claim id="CLM-00002" num="00002">
<claim-text>2. A hub-bearing assembly according to <claim-ref idref="CLM-00001">claim 1</cla
located on a substantially annular surface of the flange, in an axially external position.</
</claim>
<claim id="CLM-00003" num="00003">
<claim-text>3. A hub-bearing assembly according to <claim-ref idref="CLM-00002">claim 2</cla
annular groove, wherein the annular groove is located along the substantially annular surfac

```

Figure 20. Claims part of XML-based Patent