ภาคผนวก

ภาคผนวก ก
ตัวอย่างการเตรียมคลังข้อมูล

1384 *J. Nat. Prod.* **2003,** *66,* 1384−1387

New Antioxidant *C*-Glucosylxanthones from the Stems of *Arrabidaea samydoides*

Patrícia Mendonça Pauletti,† Ian Castro-Gamboa,† Dulce Helena Siqueira Silva,† Maria Claudia Marx Young,‡ Daniela Maria Tomazela,§ Marcos Nogueira Eberlin,§ and Vanderlan da Silva Bolzani*,†

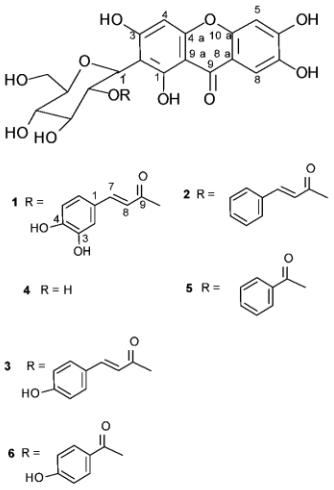
NuBBE- Núcleo de Biossíntese, Bioensaios e Ecofisiologia de Produtos Naturais, Instituto de Química, Universidade Estadual Paulista, UNESP, CP 355, 14801-970, Araraquara, SP, Brazil, Secção de Fisiologia e Bioquímica de Plantas, Instituto de Botânica, CP 4009, 01061-970, São Paulo, Brazil, and Thompson Mass Spectroscopy Laboratory, Instituto de Química, Universidade Estadual de Campinas, UNICAMP, CP 6154, 13083-970, Campinas, SP, Brazil

Received March 7, 2003

Three new *C*-glucosylxanthones, 2-(2′-*O*-*trans*-caffeoyl)-*C*-*β*-D-glucopyranosyl-1,3,6,7-tetrahydroxyxanthone (**1**), 2-(2′-*O*-*trans*-cinnamoyl)-*C*-*β*-D-glucopyranosyl-1,3,6,7-tetrahydroxyxanthone (**2**), and 2-(2′-*O*-*trans*-coumaroyl)-*C*-*β*-D-glucopyranosyl-1,3,6,7-tetrahydroxyxanthone (**3**), were isolated from the stems of *Arrabidaea samydoides*, in addition to three known *C*-glucosylxanthones, mangiferin (**4**), 2-(2′-*O*-benzoyl)-*C*-*β*-D-glucopyranosyl-1,3,6,7-tetrahydroxyxanthone (**5**), and muraxanthone (**6**). Their chemical structures were assigned on the basis of MS and 1D and 2D NMR experiments. Xanthones **1**−**6** showed moderate free radical scavenging activity against 1,1-diphenyl-2-picrylhydrazyl (DPPH) as well as antioxidant activity evidenced by redox properties measured on ElCD-HPLC.

As part of our bioprospecting program Biota-FAPESP (The Virtual Institute of Biodiversity), whose main goal is to discover potential antitumoral, antifungal, and antioxidant agents from plants of the Cerrado and Atlantic forest, we have screened hundreds of plants collected in the State of São Paulo. Among these, *Arrabidaea samydoides* was chosen for detailed chemical investigation due to prior antioxidant activity revealed on a TLC autographic assay sprayed with *β*-carotene solution, and to our knowledge there are no previous reports on chemical and biological studies. This species belongs to the family Bignoniaceae, which contains about 120 genera and 800 species distributed throughout tropical regions of South America and Africa.[1] Species from the genus *Arrabidaea* have been used in traditional medicine for wound asepsis and treating intestinal disorders.[2] In northeast Brazil, *Arrabidaea chica* is used in tattoos by Indians due to the pigments carajurin and carajurone.[2,3] A literature review indicated that this genus is a source of anthocyanins, flavonoids, and tannins.[3−7] The ethanolic extract from the stems showed promising antioxidant activity and led to the isolation of three new *C*-glucopyranosylxanthones, 2-(2′-*O*-*trans*-caffeoyl)-*C*-*β*-D-glucopyranosyl-1,3,6,7-tetrahydroxyxanthone (**1**), 2-(2′-*O*-*trans*-cinnamoyl)-*C*-*β*-D-glucopyranosyl-1,3,6,7-tetrahydroxyxanthone (**2**), and 2-(2′-*O*-*trans*-coumaroyl)-*C*-*β*-D-glucopyranosyl-1,3,6,7-tetrahydroxyxanthone (**3**), and the known mangiferin (**4**),[8] 2-(2′-*O*-benzoyl)-*C*-*β*-D-glucopyranosyl-1,3,6,7-tetrahydroxyxanthone (**5**),[9] and muraxanthone (**6**).[10] In this paper, we report the isolation, structure elucidation, and antioxidant properties of these *C*-glucopyranosylxanthones.

Compounds **4** and **6** were identified by comparison with previously published NMR and other physical data.[8,10] Compound **5** was described previously as a mixture of three isomers from *Hymenophyllum recurvum*.[9] Only a few ^{13}C NMR data were analyzed and discussed. In this paper we describe the complete ^{1}H, ^{13}C NMR and ES-MS/MS data for this compound.



Compound **1** was shown to have the molecular formula $C_{28}H_{23}O_{14}$ $[M - H]^-$ *m/z* 583.1008, by analysis of the negative-ion HRESIMS. The IR spectrum showed bands at 3370, 1615, and 1474 cm^{-1} accounting for hydroxyl, conjugated carbonyl, and aromatic groups, respectively. The ^{13}C NMR spectrum showed six signals for hydroxymethine carbons, suggesting the presence of a sugar moiety, and 22 signals for sp^2 carbons, which could be assigned to three aromatic rings, and also two carbonyls and one additional olefinic function. In the ^{1}H NMR spectrum (Table 1) of **1**, a caffeoyl moiety was identified by signals at *δ* 6.79 (1H, d, *J* = 2.0 Hz, H-2″), 6.58 (1H, d, *J* = 8.0 Hz, H-5″), and 6.67 (1H, br d, *J* = 8.0 Hz, H-6″),

* Author to whom correspondence should be addressed. Tel: 55(16)-2016660. Fax: 55(16)2227932. E-mail: bolzaniv@iq.unesp.br.
† Instituto de Química, Universidade Estadual Paulista−UNESP.
‡ Seção de Fisiologia e Bioquímica de Plantas, Instituto de Botânica.
§ Instituto de Química, Universidade Estadual de Campinas−UNICAMP.

10.1021/np030100t CCC: $25.00
Published on Web 09/24/2003

ตัวอย่างข้อมูลที่ทำการ annotation และนำเอา stop word ออก

<stuff_relation class="yes">Three new *C*-glucosylxanthones, 2-(2′-*O*-*trans*-caffeoyl)-*C*-*β*-D-glucopyranosyl-1,3,6,7-tetrahydroxyxanthone(**1**), 2-(2′-*O*-*trans*-cinnamoyl)-*C*-*β*-D-glucopyranosyl-1,3,6,7-tetrahydroxyxanthone 2-(2′-*O*-*trans*coumaroyl)-*C*-*β*-D-glucopyranosyl-1,3,6,7-tetrahydroxyxanthone isolated stems *Arrabidaea samydoides C*-glucosylxanthones, mangiferin 2-(2′-*O*-benzoyl)-*C*-*β*-D-glucopyranosyl-1,3,6,7-tetrahydroxyxanthone muraxanthone</stuff_relation>

<stuff_relation class="no">chemical structures assigned MS 1D 2D NMR experiments</stuff_relation>

<stuff_relation class=" no">Xanthones **1**-**6** showed moderate radical scavenging 1,1-diphenyl-2-picrylhydrazyl (DPPH) antioxidant evidenced redox properties measured ElCD-HPLC</stuff_relation>

<stuff_relation class=" no"> bioprospecting program Biota-FAPESP Virtual Institute Biodiversity goal discover potential antitumoral antifungal antioxidant agents Cerrado Atlantic forest screened hundreds collected State Sa˜o Paulo</stuff_relation>

<stuff_relation class=" no"> *Arrabidaea samydoides* chosen detailed chemical investigation prior antioxidant revealed TLC autographic assay sprayed *â*-carotene solution knowledge reports chemical biological studies </stuff_relation>

<stuff_relation class=" no"> species belongs family Bignoniaceae contains 120 genera 800 species distributed tropical regions South America Africa</stuff_relation>

<stuff_relation class=" no">Species genus *Arrabidaea* traditional medicine wound asepsis treating intestinal disorders</stuff_relation>

<stuff_relation class=" no"> northeast Brazil *Arrabidaea chica* tattoos Indians pigments carajurin carajurone </stuff_relation>

<stuff_relation class=" no"> literature review indicated genus source anthocyanins flavonoids tannins. </stuff_relation>

<stuff_relation class=" no"> ethanolic extract stems showed promising antioxidant isolation *C*-glucopyranosylxanthones 2-(2′-*O*-*trans*-caffeoyl)-*C*-*â*-Dglucopyranosyl-1,3,6,7-tetrahydroxyxanthone 2-(2′-*O*-*trans*-cinnamoyl)-*C*-*â*-D-glucopyranosyl-1,3,6,7-tetrahydroxyxanthone 2-(2′-*O*-*trans*-coumaroyl)-*C*-*â*-Dglucopyranosyl-1,3,6,7-tetrahydroxyxanthone mangiferin 2-(2′-*O*-benzoyl)-*C*-*â*-D-glucopyranosyl-1,3,6,7-tetrahydroxyxanthone muraxanthone</stuff_relation>

<stuff_relation class=" no"> paper report isolation structure elucidation antioxidant properties *C*-glucopyranosylxanthones</stuff_relation>

<stuff_relation class=" no">Compounds **4 6** identified comparison published NMR physical data</stuff_relation>

<stuff_relation class=" no">Compound **5** described mixture isomers *Hymenophyllum recurvum*</stuff_relation>

<stuff_relation class=" no">^{13}C NMR data analyzed discussed</stuff_relation>

<stuff_relation class=" no"> paper describe complete ^{1}H, ^{13}C NMR ES-MS/MS data compound</stuff_relation>

<stuff_relation class=" no">Compound **1** shown molecular formula $C_{28}H_{23}O_{14}$ [M - H]$^-$ *m*/*z* 583.1008 analysis negative-ion HRESIMS</stuff_relation>

<stuff_relation class=" no">IR spectrum showed bands 3370 1615 1474 cm^{-1} accounting hydroxyl conjugated carbonyl aromatic groups</stuff_relation>

<stuff_relation class=" no">^{13}C NMR spectrum showed signals hydroxymethine carbons suggesting presence sugar moiety 22 signals sp^2 carbons assigned aromatic rings carbonyls additional olefinic function </stuff_relation>

ภาคผนวก ข
ตัวอย่างประโยคที่เป็นความสัมพันธ์แบบสตัฟฟ์ที่สกัดได้

id	ตัวอย่างประโยคที่เป็นความสัมพันธ์แบบสตัฟฟ์ที่สกัดได้
88	*C*-glucosylxanthones, 2-(2′-O-trans-caffeoyl)-C-*O*-D-glucopyranosyl-1,3,6,7-tetrahydroxyxanthone 2-(2′-O-trans-cinnamoyl)-C-*O*-D-glucopyranosyl-1,3,6,7-tetrahydroxyxanthone 2-(2′-O-trans-coumaroyl)-C-*O*-D-glucopyranosyl-1,3,6,7-tetrahydroxyxanthone isolated stems Arrabidaea samydoides
217	5-epi-vibsanin G 18-O-methylvibsanin G vibsanin M aldovibsanin C isolated acetone extract leaves flowers Viburnum odoratissimum
318	(8R*)-8-bromo-10-epi-β-snyderol (8S*)-8-bromo-β-snyderol 5-bromo-3-(3′-hydroxy-3′-methylpent-4′-enylidene)-2,4,4-trimethylcyclohexanone isolated chloroform-methanol extract Laurencia obtuse
811	jujubogenin 3-O-R-L-arabinofuranosyl(1f2)-[β-D-glucopyra-nosyl(1f6) *O* -D-glucopyranosyl(1f3)]-R-L-arabinopyranoside jujubogenin 3-O-R-L-arabinofuranosyl-(1f2)-{6-O-[3-hydroxy-3-methylglutaryl]- *O* -D-glucopyranosyl(1f3)}-R-L-arabinopyranoside *O*-hydroxylup-20(29)-en-27,28-dioic acid 28-O-β-D-glucopyranosyl(1f2)-[*O*-D-xylopyra-nosyl(1f3)]- *O*-D-xylopyranosyl(1f2)-β-D-glucopyranoside ester jujubogenin 3-O-R-L-arabinofuranosyl(1f2)-[β-D-glucopyranosyl(1f3)]-R-L-arabinopyranoside *O*-hydroxylup-20(29)-ene-27,28-dioic acid isolated methanol extract stems Anomospermum grandifolium.
878	5R,9R,10β,13R-tetraacetoxy-14β-O-(β-D-glucopyranosyl)taxa-4(20),11-diene 1β,2R,9R,10β-tetrahydroxy-5R-cinnamoyoxytaxa-4(20),11-dien-13-one (2), 2R,9R,10β-trihydroxy-5R-cin-namoyoxytaxa-4(20),11-dien-13-one 9R-acetoxy-2R,10β-dihydroxy-5R-cinnamoyoxytaxa-4(20),11-dien-13-one 2R,10β-diacetoxy-1β,9R-dihydroxy-5R-cinnamoyoxy-3,11-cyclotaxa-4(20)-dien-13-one identified Taxus baccata
1188	wilfordine alatamine wilforidine alatusinine euonine euonymine ebenifoline forrestine mayteine 4-hydroxy-7-epi-chuchuhuanine isolated leaves Maytenus chiapensis.
1364	Asparagus cochichinensis led isolation asparacoside asparacosins A 3″-methoxyasparenydiol 3′-hydroxy-4′-methoxy-4′-dehydroxynyasol asparenydiol nyasol 3″-methoxynyasol 1,3-bis-di-p-hydroxyphenyl-4-penten-1-one trans-coniferyl alcohol

1734	C-glycosides 2″-O-(2‴-methylbutyryl)isoswertisin 3″-O-(2‴-methylbutyryl)-isoswertisin 2″-O-(2‴-methylbutyryl)vitexin 2″-O-(2‴-methylbutyryl)orientin 2″-O-(3‴,4‴-dimethoxybenzoyl)vitexin 2″-O-(3‴,4‴-dimethoxybenzoyl)orientin isoswertisin isolated flowers Trollius ledebouri.
2163	2,3,6,8-tetrahydroxy-1-(3-methylbut-2-enyl)-5-(2-methylbut-3-en-2-yl)-9H-xanthen-9-one isolated root bark Cudrania
2205	5R,7R,10βH-3-patchoulen-2-one 5R,7R10βH-4(14)-patchoulen-2R-ol 9R,10β-dihydroxy-2β,4β-peroxy-1R,5β,7RH-guaiane isolated aerial parts Croton arboreous
2762	10,11-dimethoxynareline alstohentine alstomicine 16-hydroxyalstonisine 16-hydroxyalstonal 16-hydroxy-N(4)-demethylalstophyllal oxindole alstophyllal 6-oxoalstophylline 6-oxoalstophyllal ones obtained leaf extract Malayan Alstonia macrophylla.
2772	14β-benzoyloxybaccatin IV 14β-benzoyloxy-13-deacetylbaccatin IV 14β-benzoyloxy-2-deacetylbaccatin VI isolated leaves stems Taxus chinensis.
3846	4β-hydroxy-19-normanoyl oxide 4R-hydroxy-18-normanoyl oxide 18-O-R-L-arabinopyranosylmanoyl oxide jhanol 18-hydroxy-13-epi-manoyl oxide isolated constituents Argentine collection Grindelia scorzonerifolia

ภาคผนวก ค

บทความการประชุม 7th International Conference on Computer Sciences and Convergence Information Technology (ICCIT2012) ณ กรุงโซล สาธารณรัฐเกาหลีใต้

Automatic Stuff Relation Extraction from Scientific Documents for Natural Product Ontology Construction

Suriyasak Lertsakunsomboon
Search Engines and Intelligent Information Systems Research Laboratory
Graduate Program in Web Engineering
Faculty of Information Technology
Dhurakij Pundit University
Bangkok, Thailand
535159090014@mydpu.net

Chaveevan Pechsiri
Search Engines and Intelligent Information Systems Research Laboratory
Graduate Program in Web Engineering
Faculty of Information Technology
Dhurakij Pundit University
Bangkok, Thailand
chaveevan.pec@dpu.ac.th

***Abstract*— To extract Part-Whole relations, especially the stuff relation, from unstructured textual data is the challenging work. This paper presents how to automatically extract the stuff relation from technical documents on the Web for supporting chemical industries. The research extracts the stuff relation without applying POS (Part-of-Speech) annotation. There are three problems of extracting the stuff relation: a) the identification of stuff relation without POS annotation problem, b) the chemical-formula-embedded name entity determination problem and c) the genus-species name entity determination problem. We propose using Naive Bayes to learn the stuff relation. The results from our proposed methodology are 87% precision and 61% recall.**

Keywords-stuff relation; scientific name entity; chemical name entity

I. INTRODUCTION

Through out history, there are consistent amount of interests to extract chemicals of natural products for using in different areas of the pharmaceutical industry [1], the alternative energy research [2], and the commercial biorefinery system [3]. Therefore, several natural product researches work on analysis of natural substances of land and sea and of plants, microbes and animals where the research results bring significant benefits to the industries, especially the pharmaceutical industry. For example, many natural product researches focus on identifying novel bioactive constituents from mushrooms and medicinal plant. Recent studies identified antibacterial and cytotoxic mushroom species (where cytotoxic mushroom uses as anti-cancer medicines), as well as novel compounds from Bangladeshi medicinal plants. Since numerous scientific literatures exist on natural substances, the allocation of large amounts of time and resources are required by industries to seek source organism alternatives for a specific constituent. Therefore this research proposes to automatically extract the Part-Whole relation (i.e. "X is the component of Y") from the research documents to reveal the source organism substances of the natural product. According to M.E. Winston et. al. 1987, the Part-Whole relation can be classified into six types: Component-Integral object (wheel-car), Member-Collection (soldier-army), Portion-Mass (meter-kilometer), Stuff-Object (alcohol-wine), Feature-Activity (paying-shopping), and Place-Area (oasis-desert). Therefore, this research aims to determine and extract the Part-Whole relation, especially Stuff-Object type, from the documents. Unlike the Component-Integral relation, the Stuff-Object relation (or Stuff relation) refers to a relation where the stuff cannot be physically separated from the object without altering its identity [4]. The Stuff relation is required for automatically constructing the natural product Ontology used to represent all natural product knowledge.

There are several researches work on the relation extraction [5][6][7][8], but literature on the Part-Whole (or Part-of) relation extraction is still lacking. Most of these researches worked on the text file involved with the linguistic patterns or the linguistic rule bases at the phrase level (e.g. "the oil and vinegar salad dressing") or the sentences level (e.g. "A vinaigrette-type salad dressing in United States and Canadian cuisine consists of water, vinegar or lemon juice, vegetable oil, chopped bell peppers corn syrup, and a blend of various herbs and spices."). Our research concerns only the sentence level since the Stuff relation expression on our corpus of the natural product research papers mostly occurs at the sentence level. In addition, our corpus contains several characteristics differ from other general corpora, especially the name entity (especially IUPAC nomenclature or chemical name entity) such as the chemical-formula-embedded name entity as shown the following with the underline.

"*Seven new lanostane-type triterpenes, hypo-crellols A−G (1−7), and six new hopane-type triterpenes, 7β,15α-dihydroxy-22**(29)**-hopene (8), 3β,7β-dihydroxy-22(29)-hopene (9), 3β-acetoxy-15α-hydroxy-22(29)-hopene (10), 3β,7β,15α,22-tetrahydroxyhopane (11),*

3β-acetoxy- 7β,15α,22-trihydroxyhopane (12), and 7β,15α,22-trihydroxy-hopane (13), were isolated from the scale insect pathogenic fungus Hypocrella sp. BCC 14524."

Where each number in the parenthesis except **(29)** stands for the identification number of the element in front of the parenthesis, for example:

"1−7" (1 to 7) stand for "*lanostane-type triterpenes, hypocrellols A−G*"

"8" stands for "*hopane-type triterpenes, 7β,15α-dihydroxy-22(29)-hopene*"

Another problem is the chemical name entity contains several commas whose functions differ from the general comma function separating word in a series (www.brighthubeducation.com). For example:

"*… … …, 3,4-dihydroxy-9-methoxypterocarpan (vesticarpan) (5), 2′,4,4′-trihydroxychalcone (isoliquiritigenin), and 7,4′-dihydroxyflavanone (liquiritigenin) (6), were isolated from the heartwood of Platymiscium floribundum.*"

where "*2′,4,4′-trihydroxychalcone*" is one word. Whereas "*dogs*" / "*cats*" / "*mice*" is a one word in the following sentence.

"*Dogs, cats, and mice are mammals.*"

All of these characteristics are involved in the three main problems; first is how to identify the Stuff relation. Second is how to determine the chemical formula name entity. Third is how to determine the organism (species) name entity. From all of these problems, we propose applying the Naïve Bayes machine learning technique to learn the stuff relation from a sentence that contains the interesting word set {"*obtain*", "*isolate*", "*extract*", …}(from corpus behavior study) along with the stuff word set {"*7,4′-dihydroxy-3′-methoxyisoflavone*", "*isoliquiritigenin*", "*(R)-4″-methoxydalbergione*", …}(from NBIC-PubChem database) and the object word set { "*Dalbergia louvelii*", "*Dendrolobium lanceolatum*", … }(from NBIC-Taxonomy database).

In section II, related work is summarized. Problems in extracting the Stuff relation from the published research papers is described in section III and in section IV is purposed our framework for the Stuff relation extraction. In section V, we evaluate and conclude our proposed model.

II. Related Works

There are some previous researches including [5][6][7][8] working on the relation extracting from texts as described in the following.

R. Girju et. al.(2003)'s work [5] is to present a ID3 (C4.5) learning technique for learning semantic constraints based on the lexico syntactic pattern , NP1 verb NP2 where NP1 and NP2 are noun phrases, to detect Part-Whole relations (meronymy) from the LA Times articles of TREC 9 text collections. They also stated that the Part-Whole relations in WordNet were classified into three basic types: Member-of (e.g., UK IS-MEMBER-OF NATO), Stuff-of (e.g., carbon IS-STUFF-OF coal), and Part-of (e.g., leg IS-PART-OF table). According to their proposed model based on sentence level, the Part Whole relations are detected with an accuracy of 83% precision 98% recall from 10000 sentences.

In 2006, P. Pantel and M. Pennacchiotti's work [6] is to present the weakly-supervised algorithm, named Espresso, using the generic patterns to extract the semantic relations, especially the Is-a relation and the Part-of relation. The Espresso was applied to the very large corpora downloaded from webs with the generic pattern which was determined the pattern and instance reliability. The reliable patterns resulted in having high precision but often very low recall (e.g., "X consists of Y" for the Part-of relation). Their experimental results showed that their generic patterns substantially increased system recall with small effect on overall precision. The results of their system performance using the CHEM corpus with the Part-of relation is 51% precision and 46 relative recall whereas using TREC-9 with the Part-of relation is 70% precision 577 relative recall.

In 2007. K.Fundel et. al. [7] developed RelEx, an approach for relation extraction from free text, especially biomedical publications from million MEDLINE abstracts. RelEx was based on natural language preprocessing producing dependency parse trees dealing with gene and protein relations. The RelEx model involved the detection of co-occurrences of entities within sentences or abstracts and uses a small set of simple rules, applied for part-of-speech-tagging, noun-phrase-chunking and dependency. Their Relation Extraction is based on three linguistic rules frequently used in English language for describing relations where one rule is based on the sentence level and the other two rules are based on the phrase level as follow:

(1) effector-relation-effectee ('α activates β')

(2) relation-of-effectee-by-effector ('Activation of α by β')

(3) relation-between-effector-and-effectee('Interaction between α and β').

Finally, RelEx is estimated performance of both 80% precision and 80% recall.

G. I. Brown (2011) [8] presented an analysis of a relation extraction system using a support vector machine (SVM) classifier to the J.D. Power and Associates Sentiment Corpus separated into three style documents: professionally written reviews, blog reviews, and social networking reviews. The SVM features includes the word features involved the head noun of the phrase existing the least deep in the dependency parse tree, the entity types, and the token class. However, his research aims to study how the extraction system works on different styles of documents. The results of the Part-of relation extraction are 46%precision on average and 33% recall on average. Then, [8] concluded that the relation extraction task was being negatively impacted by the relation

classification itself and the poor tokenization or parsing of the documents.

However, all of the relation extraction techniques from [5][6][7][8] cannot be applied to our research. The complicated chemical-formula name entity and the organism name entity limit the use of POS standard tools (http://open.xerox.com/Services/fst-nlp-tools/Consume/181). Therefore, we apply Naïve Bayes to learn the Stuff relation from a sentence that contains the interesting-stem word set along with the Stuff word set and the object word set.

III. Problem of Stuff Relation Extraction

There are three main problems: how to identify the Stuff relation, how to determine the chemical-formula-embedded name boundary, and how to determine the scientific name boundary.

A. How to identify Stuff Relation without POS annotation

In order to identify the Stuff relation without the POS annotation, we propose using the A-B-C or C-B-A sequence (where A is the Stuff word set, B is the interesting word set, and C is the object word set) occurring within one sentence to identify the Stuff relation using Naïve Bayes. A and C are obtained from NBIC-PubChem and NBIC-Taxonomydatabase (http://pubchem.ncbi.nlm.nih.gov) respectively (see Fig. 1.).

> "… Four new flavonoids (1-4), along with 13 known compounds, were isolated from the heartwood of *Dalbergia louvelii*. ..."

Fig. 1. Example of the linguistic expression in scientific documents.

where A is "flavonoids (1-4)", B is "isolate", and C is "*Dalbergia lovelii*".

B. How to determine the chemical-formula-embedded name boundary

The chemical formula is very complex word (as shown in TABLE I.) where the chemical names are embedded within another chemical name (see section I). Therefore we used NBIC-PubChem to determine the chemical name.

TABLE I. List of Chemical Formula Example.

Chemical Formula
Example 1 : (R)-4″-methoxydalbergione
Example 2 : 3-(2,4-dihydroxy-5-methoxy)phenyl-7-hydroxycoumarin
Example 3 : (R)-4″-methoxydalbergione
Example 4 : (7S,8R,1′S,5′S,6′R)-$\Delta^{2',8'}$-5′,6′-dihydroxy-3′-methoxy-3,4-methylenedioxy-4′-oxo-8.1′,7.5′-neolignan
Example 5 : 2,4-dimethoxy- 5,6-methylenedioxy-1-(2-propenyl)benzene
Example 6 : 2′-hydroxy- 6,4′,6″,4‴-tetramethoxy-[7-O-7″]-bisisoflavone

Also, the chemical formula often contains "," as separating word (see section I). We propose using NBIC-PubChem database to determine the chemical formula name entity.

C. How to determine the scientific (species and genus) name boundary

Often the species and genus names of species are written in multiple word italic form (as shown in TABLE II.). However, when the surrounding texts are also in italics, the identification of the scientific species names become a challenging task.

TABLE II. List of Scientific Name Example.

scientific name
Example 1 : *Dalbergia louvelii*
Example 2 : *Dendrolobium lanceolatum*
Example 3 : *Hypericum perforatum*
Example 4 : *Platymiscium floribundum*
Example 5 : *Dalbergia candenatensis*

Therefore we propose using NBIC-Taxonomy database to solve this problem.

IV. Methodology

There are 4 major steps including corpus preparation, name entity determination, Stuff relation learning, and Stuff relation extraction as shown in Fig. 2.

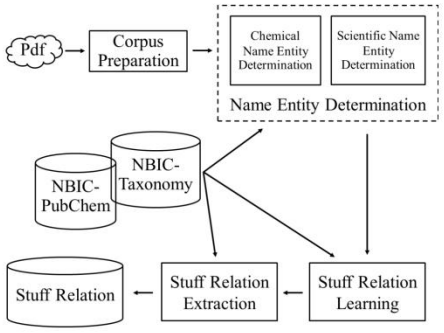


Fig. 2. System architecture.

A. Corpus Preparation

In the corpus preparation step, English scientific abstracts and introductions in chemical related areas are downloaded from online journals (20,000 sentences from 500 scientific documents in pdf format). The pdf documents are converted to text documents using PDFTextStream (http://snowtide.com). The corpus is separated into 2 parts; one part is 15,000 sentences for learning the stuff relation and the other part of 5,000 sentences for stuff relation extraction. The sentences with stuff relation are manually annotated as "yes" class and the others are "no" class (see Fig. 3.). Then the stop words are filtered out while all symbols , e.g. "," "-" … etc., still exist in the corpus.

<stuff_relation class="yes">Four new flavonoids (1-4), along with 13 known compounds, were isolated from the heartwood of *Dalbergia louvelii* by following their potential to inhibit in vitro the growth of *Plasmodium falciparum*.</ stuff_relation>
<stuff_relation class="no">Of these, the ethyl acetate extract obtained from the heartwood of *Dalbergia louvelii* R. Viguier (Fabaceae).</stuff_relation>
<stuff_relation class="yes">Although several isoflavonoids have been obtained from roots of *P. floribundum*, none of the abovementioned compounds have been isolated previously from this species.</stuff_relation>
<stuff_relation class="no">The cytotoxicity of the isolates obtained herein from *P. floribundum* has been evaluated against a small panel of cancer cell lines.</stuff_relation>

Fig. 3. Examples of sentences annotation.

B. Scientific Name Entity Determination

1) Chemical Name Entity Determination: There are 2 steps involved: the translation of numeric representation of chemicals and the identification of chemical name entities.

a) Translating Numeric Representation of chemicals: The following observed rule are used to translate the numeric representations of chemicals.

Rule:

if((Cname is the first occurrence)∧ (Cname(num))then
num is the numeric representation of Cname

Where Cname = chemical name on a scientific document
num = the integer

b) Chemical Name Entity Identification: The results from a) coupled with the adjacent surrounding word are compared to the NBIC-PubChem.

2) Genus-species Name Entity Determination: Using NBIC-Taxonomy database solves the Genus-species Name Entity.

C. Stuff Relation Learning

In the learning step, b (where b ∈ B and B is obtained from the corpus behavior studied) is used to anchor sentences. Sentences where the left hand side of the anchor contains a (where a ∈ A) or c (where c ∈ C) and the right hand side of the anchor contains c or a are collected. Then the frequency of a, and c with class "yes" and class "no" are determined for each sentences (see TABLE III.)

TABLE III. FREQUENCY OF A AND C WITH CLASS "YES"/"NO"

A	Class=yes	Class=no
3,10-dihydroxy-9-methoxypterocarpan	0.05882353	0.03571429
ethyl acetate	0.05882353	0.17857143
dibenzocycloheptene	0.3921569	0.03571429
…	…	…
C	**Class=yes**	**Class=no**
Dendrolobium	0.06976744	0.05
Platymiscium	0.34883721	0.25
Dalbergia	0.06976744	0.25
…	…	…

D. Stuff Relation Extraction

In order to start the Stuff relation extraction process, Naive Bayes Classifier[9] shown in equation (1) is applied. The class "yes" means Stuff relation, as shown in Fig. 4.

$$\begin{aligned} Stuff\ \mathrm{Re}lationClass &= \arg\max_{class \in Class} P(class \mid a, c) \\ &= \arg\max_{class \in Class} P(a \mid class) P(c \mid class) P(class) \end{aligned} \quad (1)$$

$$where\ \ Class = \{"yes", "no"\}$$
$$a \in A \quad (A\ is\ a\ Stuff\ word\ set)$$
$$c \in C \quad (C\ is\ an\ object\ word\ set)$$

```
L is a list of sentence.
W is a set of word in each sentence.
WS is a sequence of n words where n = 1, 2, 3, … .
C is natural-product-compounds concept set.
S is the natural-sources concept set.
A is the marker set.
1   i ← 0, R ← ϕ
2   while (i ≤ length[L]) do
3   Begin
4     j ← 0, len ← length[Wi]
5     while(wij ∈ A) do
6       for (k ← 0 to j)
7         if (wsik ∈ C) then
8           for (m ← j+1 to len)
9             if ((wsim ∈ S) ∧ (StuffRelationClass
              == "yes")then
10              R = R ∪ {(wsik, wsim)}
11            end if
12          next
13        else if (wsik ∈ S) then
14          for (m ← j+1 to len)
15            if ((wsim ∈ C) ∧ (StuffRelationClass
              == "yes")) then
16              R = R ∪ {(wsim, wsik)}
17            end if
18          next
19        endif
20      next
21  End
22  return R
```

Fig. 4. Stuff Relation Extraction Algorithm.

V. EVALUATION AND CONCLUSION

The English corpora of the technical documents in chemistry domain are used to evaluate the proposed stuff relation extraction algorithm consisting of about 5,000 sentences. The evaluation of the Stuff Relation extraction performance in research is expressed in terms of the precision and the recall as shown below, where R is the stuff relation:

$$\text{Precision} = \frac{\#\text{ of samples correctly extracted as R}}{\#\text{ of all samples output as being R}} \quad (2)$$

$$\text{Recall} = \frac{\#\text{ of samples correctly extracted as R}}{\#\text{ of all samples holding the target relation R}} \quad (3)$$

The results of precision and recall are evaluated by three expert judgments with max win voting. The precision of the extracted Stuff relation is 87% and 61% recall. The reason of the recall limited to 61% is misplaced-B problem where B is before or after A and C. These problems are a subject of further studies. Finally, these extracted Stuff relations are beneficial for natural product chemical ontology construction (see Fig. 5.) which will bring significant benefits to the cosmetics, pharmaceuticals and other chemicals industries.

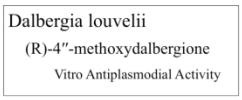


Fig. 5. Example of Natural Product Chemical Ontology.

REFERENCES

[1] Bernard Munos, "Lessons from 60 years of pharmaceutical innovation," Nature Reviews in Drug Discovery, vol. 8, pp. 959-968, December 2009.

[2] Yusuf Chisti, "Biodiesel from microalgae," Biotechnology Advances, vol. 25, pp. 294–306, February 2007.

[3] Gail Taylor, "Biofuels and the biorefinery concept," Energy Policy, vol. 36, pp. 4406–4409, 2008.

[4] Morton E. Winston, Roger Chaffin, Douglas Herrmann, "A taxonomy of part-whole relations," Cognitive Science, vol. 11, pp. 417-444, 1987.

[5] Roxana Girju, Adriana Badulescu, Dan Moldovan, "Learning semantic constraints for the automatic discovery of part-whole relations," In Proceedings of HLT/NAACL-03, vol. 1, pp. 1-8, 2003.

[6] Patrick Pantel, Marco Pennacchiotti, "Espresso: Leveraging generic patterns for automatically harvesting semantic relations," In Proceedings of COLING/ACL-06, pp. 113-120, 2006.

[7] Katrin Fundel, "Text Mining and Gene Expression Analysis Towards Combined Interpretation of High Throughput Data," München, 2007

[8] Gregory Ichneumon Brown, "An Error Analysis of Relation Extraction in Social Media Documents," In Proceedings of ACL-HLT-49, pp. 64-68, June 2011.

[9] Tom Mitchell, Machine Learning. Singapore: The McGraw Hill Companies Inc. and MIT Press, 1997.