Image tasks @ CLEF-IP 2012

Mihai Lupu, Florina Piroi, Allan Hanbury, Igor Filippov, Alan Sexton

Overview

- Chemical structure recognition results
- Flowchart recognition results
- Next steps

Chemical Structure Evaluation

Mihai Lupu Igor Filippov Alan Sexton

Chemical structure recognition

a group such as tationity) or ${\rm COMR}^4$ in which ${\mathbb R}^4$ is hydrogen, lower allighter helically):

[0034] is a set p are interpendently 0 or 1, and j in

From alley, allerati, hais, traicolleyt, alleray, S-alley, ME⁴⁴-aliqi, aryl or herecorryl, which are preferably single rings that compile 4 to 7, more preferably 5 or 6, transferab in the

cycloalleysyl, and is preferably hydrogen, lower alkyl, h alizery and lower halostkyl, and

-CHACE TOOL

2, preferably this

R⁴ is hydrogen, hower alkyl or helenikyl;

0043] R⁷ is selected from bydrogen, silvel or balcolicy!

[9844] R* is induced, an induction -flower all of ICO.R*

[3048] R^{*} is --CN, -CO, R¹⁰, -CH, OH, or earbarrent

[0048] R¹² and R¹³ are independently hydro -CO₂R¹⁰, --CN, aryl, lower alkyl, heteroaryl lower i or --NHCOMPAL provided that one of R¹⁴ and R¹⁵

architerri or -- CO.H.

or ballenillered 100411 R²⁰ and R²⁷ are such independently selected from

0050 R¹⁵, R¹⁶ and R¹⁶ are independently selected from

[2004] K. and K. are said inseptiming second to hydrogen, alled, allowed, alleged, ard, holoalist, allefaryt heurocyclyl, arylallyd, arylallway, allows, aryloap secondleyl, systemized and cycloallyney).

is hydrogen, alleyl, slower alkylications, arg-

[0046] R^{ar} is --CO.H or carbourghamit.

[0047] R¹¹ is balanges, slipil or architect

[9657] X. Y. Ar¹ and Ac² are predictional at

(00M) Preferred community for our in have formula (I) and (II), but with

Ar in phonel, X in Hork, Yi

DEAD A M

dd, sa Mol

100351 R², R⁴ and R²⁰ are each independently a

pulkin My O.

(associated)

US 2001/0014694 A1

cycloshyl, cycloshuryl, cycloshuryl, OH, R⁴¹, C(OR⁴⁰, COR⁴¹, SH, SHO, R⁴¹ in which is 50-2, INOH, ICH, JH, C(E), JR⁴² in which is is 1-5, NR⁴²R⁴¹, OR⁴², R⁴¹NCOR⁴² and R⁴¹NHO, R⁴²,

[9026] R^m is soluted from among hydrogen, aligh, alk mp1, alignpt, argt, alighest, holereceptyt, argtaligh cycloalityi, cycloalita nyi or cycloalitynyi, where R²⁰ is perforably alityi or aryi; and [9027] R⁺⁺ is solution from among hydrogen, allest, alle

ees), elicropi, eeri, aliatard, alieney, acriene, heineinevisi arybileyi, arybileosy, eprioaikyi, eprioaikanyi, eprioaiky

are represented by the state of the state o with an acidic gamp, particularly a mathemyl gamp or an incohere the real

[8629] It is noted that the compounds for compositions do not have the formula CIVE



shish IC is lower alkel, COOR, COONE R. [1906] in stands K is lower adapt C1004, (1)(1908; $K_{0,2}$, where $R_{0,1}$ is belowing as $C_{1,2}$ -adapt, $R_{1,1}$, $C_{1,2}$ -adapt, $M_{1,2}$, $C_{1,2}$ -adapt, $M_{1,2}$, matherly, epsemberly, or $R_{1,2}$ and $R_{1,2}$ regardler from $=(O14)_{A_{1,2}}$ -scheme as 1 to 5.1 m for performal comproseds $R_{1,2}$ is also $(C_{1,2}-add)$ (CO006). Compose alls of formula DV_{1} forwards, any be used in the contlasts.

Most preferred among the compounds that are used methods are those of formula (1) that have formula



given a full-text patent PDF and TIFF

a group such as tanzacily) or COOMR⁴ in which R⁴ is hydrogen, lower allighter helpeliky);

Aug. 16, 2001

[00043] , a and p are independently 0 or 1, and (in prefer by 0; [2015] K. R. and K.¹⁰ are each independently whereas from alleyt allerest, built indexists, allows, S. alkyl, SR¹⁰, alkyl, argl or horizonyl, which are penfittably single rings that coming in C. 7, more penfittably as in the.

[9834] R²⁰ is selected from bridrogen, allot, allocat, allo (66) K. 'to set of our top-to-gen, staty, starty, interpl. tab-i. arXi, binardidj, starty and the set of the set of the set interplanet, stephilacep, opticality, opticalizing and challengel, and is prefamily hydrogen, lower sikyl, hear imp and hence halasticpt, and

[9837] X, Y, M² and he² are as defined above.

(9034) Performi composeds for one in the compositions have formula (I) and (II), but with the provise that the compounds its not have formula (IV). In particular, whe At is planyl, X is 0 or S, Y is 0 or S, and At' and At' are waveletized or substituted with halogon or lower alkyl, iten R¹ is not COOH or COOPE R. . where R. is bullet silkyl, R₁ is C_{1,2} alkyl, OH, maltery, cyanomethyl and R₂ together form —(CH₄)₂—, where a is 1 in t R and R, logs

cyl, thiocarbameyi, or a citragen-containing ring, where

100400 Ais substant from among CO.R⁴, carborrelic acid do, aller Rhinig acid. all relations used, aller binishedd, willow sold, sufficie sold, phosphonic, sufficiently, add, sufficiently, and teaming, silvelly decraming and hydrackle, amile, hydroxyl, hydrogen

[9841] q is 0 is 12, preferably 0 to 0;

[9842] R⁴ is hydrogen, hear aligh or holosikyl [9843] R7 is selected from hydrogen, alkyl or halcolcyl;

[9844] R' is hydrogen, a tylish ylor-thour alkyli CO,R". $[3043] \quad \mathbb{R}^n \text{ is } -CN, -CO, \mathbb{R}^n, -CH, OH, \text{ or surface pl};$ [9946] R¹⁰ is -CO. H or exterrupturely

[9647] R¹¹ is below a distorteriolistic [0048] R¹² and R¹⁵ are independently hydrogen, —CO₂R¹⁰, —CN, argl, lower aligh hearcarritower aligh or —NHCO(repl., provided that use of R¹⁶ and R¹⁶ is

9848] R¹¹ is hydrogen, skyl, -flower alkyljearbory, ary-idlawel, hotowaryladlawel or --C0.16. 0050] K¹⁷, K¹⁴ and K¹⁴ are independently selected from redrogen, alleyl or baleniket.

[0051] R²⁰ and R²⁰ are such independently selected from by trying a set of the [0852] i. o and p are each independently 0 or 1

US 2001/0014694 A1

cyclosifyi, cyclosharyi, cyclosheryi, GH, R⁴⁰, COOR⁴⁰, COOR⁴⁰, SH, SIO, R⁴⁰ in which a is 0.2, BHOH, IGEL, IGEL, IGEL, ISABA a is 1.4, NR⁴⁰R⁴¹, OR⁴⁰, R⁴⁰NCOR⁴⁰ and R⁴⁰NSO, R⁴⁰;

[8026] R¹⁰ is soluted from among hydrogen, alityl, sik-melt, allipseyl, asyl, alliplayit, balancepulyt, asylatikel, cycloalized, epchoalias nyl or cycloalizenyl, where R⁴⁰ is performed with a state or anyl, and

[9027] R²¹ is salested from among bydrogen, slipil, alkepi, diqupi, ord, diplayi, along, aylon, horacychi aylaliyi, aylalion, sydoaliyi, sydoalianyi, sydoaliy $P(\mathbf{r}) = P(\mathbf{r}) + \frac{1}{2} \left[\frac{1}{2} \left[$

has, Ar³ is preferably selected from arrang pherel, pyrin (b) praticel, prekateri, prility, ocardyl, iestatski rak iniskecist groups. Ar³ is more prelicably physic, ist, preting dyl or presented and is preferably substituted with an acidic group, particularly a carbonyd 0079] It is noted that the corm

do not have the formula CIV

ber R_n is by-drogen or C₁₁₀-tileyt, B_n, theory, space-methyl, or R_n and S (CH₂)_n..., where n is 1 to 0. In the pro-is also (C₁₁₀-tileyt)COOM. Compare is also (C₁₁₀-tileyt)COOM. Compare 001] Max puls



9032) where [803] B¹ is hydrogen at =0.034,1,...,A in which a is 0 is b, parkeddy 0 to 3, and A is an actile group. A is preferably (8852) 1, o and p are each independently 0 or 1

identify the bounding boxes of chemical compounds









crop out the individual chemical images

generate chemical structure information

-6.4573 -0.4004 0.0000 C

-6.4573 -1.000

-1.1548

-7.0410

-3.0618

.

0.3099 -1.0108

1.0404 -7.4471

3.1342 -1.000

3.1342

1.8404

6.4366

8.7771 -0.4004

8.7775 -1.0008

7.0776

8.3333

1.0774

1.0000

4.1010 2.4572

4.8710 2.4572

10.11

.....

.

11 12 4 5 0 5 0

Chemical structure recognition

- Good results for both tasks
 - Segmentation

Tolerance	Precision	Recall	F ₁
0	0.70803	0.68622	0.69696
10	0.79311	0.76868	0.78070
20	0.82071	0.79543	0.80787
40	0.86696	0.84025	0.85340
55	0.88694	0.85962	0.87307

- Recognition

	Automatic Set			Manual Set			Total		
	#Structures	Recalled	%	#Structures	Recalled	%	#Structures	Recalled	%
saic	865	761	88%	95	38	40%	960	799	83%
uob-1	865	832	96%	95	44	46%	960	876	91%
uob-2	865	821	95%	95	56	59%	960	877	91%
uob-3	865	821	95%	95	44	46%	960	865	90%
uob-4	865	832	96%	95	54	57%	960	886	92%

Chemical structure recognition

Good results for both tasks

865

uob-4

832

96%

- Caveat The test structures have been pre-selected to have a INCHI representation (i.e. no Markush, no 'fancy stuff') said uob uob-2 9591% 865821 95% 5659%960877 uob-3 865 821 95% 9546%960865 90% 44

95

54

57%

960

886

92%

Flowchart Recognition Evaluation

Mihai Lupu Florina Piroi Allan Hanbury

• Most common sub-graph



Node type match



Text recognition (Edit Distance) - average

• Text recognition (Edit Distance) - sum

• Text recognition (Edit Distance) - normalized

Caveat

- 44 of 100 topics evaluated
 - Only those where a score was computed for all participating runs
 - 'selected' by the algorithm (by not finishing in reasonable time or crashing)

	Number of nodes					
	Flowcharts in results	Flowcharts not in results				
Min	6	10				
Max	21	51				
Median	13	22				
Average	12.41	24.98				
std.dev	3.62	8.9				

Examples of finished evaluations

[FIG. 2]

FIG.2

À

OPERATION

OPERATION AT ACD CONTROLLER

Unfinished evaluations

Measures

- Idea / Motivation
 - The process depicted is important
 - Links between nodes
 - Graph structure
 - » Most Common Sub-graph
 - Node types are less important
 - Only evaluate after structure recognition
 - Text is important
 - Provided that the structure is recognized

Measures (cont.)

- Most common sub-graph (MCS)
 - It's the largest sub-graph common to all graphs in a set of graphs
 - McGregor algorithm
 - Backtracking
 - High computational costs
 - Modified to find all variants of the most common subgraph
 - Because (we think) filtering on node type would be too restrictive
 - Even more complex

Measures (cont.)

- Node type match & Edit distance
 - Taken separately as the best of all different variants of largest common sub-graphs
 - E.g. if for topic X, 5 different ways to match nodes of run Y were found (all having a score of 0.7 in the MCS), compute the node type match & edit distance for each
 - Node-type match : 0.5 0.5 0.9 0.9 0.5
 - Edit distance: 100 100 90 120 90
 - Then the result scores are:
 - Node match: 0.9
 - Edit distance: 90

Edit Distance

- Smaller is better
 - Smaller?
 - no nodes were returned
 - Actually good match
 - Only comparable in relation with MCS

What's up next?

• But first...

Information Retrieval Special Issue on IR in the Intellectual Property Domain

- Goal: present cutting-edge research results on open topics related to IR in the Intellectual Property domain in order to advance the current state-of-the-art
- Submissions encouraged to make use of evaluation campaign datasets: CLEF-IP 2011/2012, TREC-CHEM 2011 and NTCIR
- Submissions due: **15 March 2013**
- More info: <u>https://sites.google.com/site/sipatentir</u>

What's up next?

• Overview and plans

Image retrieval in patents

- Must be approached by image type
- Step 1: $\frac{km}{b}s+k$ - classification ms² (a) Mathematical (b) Chemical Formula (c) Genetic Sequence (d) Flow-chart -+ IRON CORE · CADWELL CC O CADWELL - 84 MAGSTIM 13.5 CM NACSTM 2-COR 5 COLS IN AR 92a ---- RON-CORE CADWELL 90a 57a 的数数数数数数数数 COLS IN SALM FIG.7 Fig. 10 (e) Graph (f) Program Listing (g) Abstract Drawing Table 1 V

(h) Symbol (i) Table

FIG.

Figure 1: Examples of types of figures in patents

Figure 1: Examples of types of figures in patents

Mathematical formulas

Chemistry

Genetic data

HindIII

-----|-----

5'-untranslatiert

Beginn Leserahmen (Signalpeptid)

Plots/Graphs

FIG.7

Code listings

402 cla	ss Hashta	able\$Checked : IDictionary {
404	void Has	htable\$Invariant() {
405	A	ASSERT([[constraint clause from Fig. 3]]);
406	}	
407	void set\$	Pre (Object key, Object val) { ASSERT(true); }
408	void set\$	Post (Object key, Object value, Object result) {
410	A	ASSERT([[ensure clause from Fig. 3]]);
412	}	
414	Object[]	keys; Object[] values;
418	Object se	et (Object key, Object value) {
420	0	Object result ;
422	F	Hashtable\$Invariant();
424	s	set\$Pre(key, value);
426	t	ry {
428		[[body of the set method from the implementation code]]
432		<return break="" end;="" result="value;" value=""></return>
434	}	catch (Exception e) {
436		result = e;
438	}	•
440	END :	
444	ŀ	Hashtable\$Invariant();
446	5	set\$Post(key, value, result);
448	i	f (result is Exception) throw result ; else return result ;
450	}}	

Drawings

Fig. 10

Characters/symbols

Tables

Table 2

Sample	Total amount of	Calcining	Firing	Mechanical	Resistance	Moisture	Glass	phase	at
No.	network-forming	temp.	temp.	strength	to chemicals	resistance	grain	bounda	ary
	oxides (ppm)	(°C)	(°C)	(Kgf/cm ²)	(%)	(%)			
2	230	970	1210	1463	-3	-5	No		
3	190	950	1170	1444	-5	-8	No		
4	630	810	1100	1559	-1	-6	No		
5	460	1000	1280	1695	-1	0	No		
6	400	840	1210	1396	-8	-8	No		
7	320	850	1230	2039	0	0	No		
		1							