



إلى القلب الدافع إلى من نور لي دربي إلى نبع الحنان الذي لاينضب إلى الحضن الذي ألجأ إليه عندما  
أتعب إلى لمسة الأمل إلى من أرى ملامح الأصالة فيها .....

إلى أبي الحبيبة

إلى من ي العمل على غرس بذور الخير في شهي وينير نبراس الأمل في روحي إلى من يسعى لأرتاح  
ويدفعني نحو النجاح إلى رمز التضحية .....

إلى أبي الغالي

أُسامَة سليق



إلى من وضعاني على درب النجاح وهيئا لي جميع سبله .. إلى معلمي الأوليين

إلى أبي وأمي ..

إلى من ساندني وكان عونا لي ..

إلى أخي ..

مررت معكم بأيام ستنظل عيناي تفياض بالدموع كلما ذكرتها ..

إلى أصدقائي ..

آنا عجميان



**بلال الهملا الشريفي**

إلى من علمني الكلمة الأولى

إلى الذي تعب وكد من أجل مستقبلني

إلى من كان مظلتي من مخاطر حياتي

إلى والدي الغالي...

إلى القلب الحنون وعشقي الأول

إلى من سهرت ليالي لتضيء دربي بالنجاح

إلى أمي الغالية...

إلى من هم أقرب إلى نفسى مني

إلى أصدقاء ماضي وحاضر ومستقبلني

إلى أخوتي وأخواتي...

إلى من ملؤوا حياتي فرحاً وسعادة

إلى من شاركوني هموي

إلى أصدقائي...



إلى كل من كان بجانبي طوال تلك السنين ..

إلى من سكنت صورهم قلبي ..

إلى من نكن لهم خالص المحبة والاحترام ..

إلى من أدين لهم بالجميل والعرفان ..

إلى بلدي الحياة والصمود ..

إلى سوريا وأهلها ..

إلى أمي ..

إلى أبي ..

إلى أخوي ..

لما موازني



# فهرس المحتويات

## Contents

الفصل الأول مدخل إلى المشروع.....	
24.....	
24.....	مقدمة 1
26.....	عرض المشكلة 2
27 .....	أهداف المشروع 3
28 .....	إدارة المشروع 4
28 .....	الموارد المتاحة 4.1
28 .....	الموارد العتادية 4.1.1
28 .....	الموارد البرمجية 4.1.2
29 .....	الموارد البشرية 4.1.3
29 .....	المهام الرئيسية 4.2
29 .....	مرحلة الدراسة النظرية 4.2.1
30 .....	مرحلة التحليل 4.2.2
30 .....	مرحلة التصميم 4.2.3
30 .....	مرحلة التحقيق 4.2.4
30.....	مرحلة الاختبارات 4.2.5

31 .....	الجدول الزمني للمهام	4.2.6
33 .....	مخططات غانت	4.3
33 .....	مخطط غانت لمرحلة الدراسة النظرية	4.3.1
33 .....	مخطط غانت لمرحلة التحليل	4.3.2
34 .....	مخطط غانت لمرحلة التصميم	4.3.3
34 .....	مخطط غانت لمرحلة التحقيق	4.3.4
35 .....	مخطط غانت لمرحلة الاختبار	4.3.5
36 .....	الباب الأول الدراسة المرجعية.....	
38 .....	الفصل الثاني مشاريع في فلترة صفحات الإنترنت .....	
40 .....	مشاريع في فلترة صفحات الإنترن.....	1
40 .....	<b>DansGuardian</b>	1.1
43 .....	<b>K9</b>	1.2
46 .....	<b>OpenDNS</b>	1.3
48 .....	<b>We-Blocker</b>	1.4
49 .....	<b>Anti-Porn</b>	1.5
51 .....	<b>Squid</b>	1.6
52 .....	<b>SquidGuard</b>	1.7
54 .....	<b>SafeSquid</b>	1.8
56 .....	<b>QuintoLabs</b>	1.9
60 .....	الفصل الثالث دراسة مرجعية عن فلترة وتصنيف البريد الإلكتروني.....	

61	ادارة Outlook للبريد الإلكتروني من خلال القواعد .....	1
62	Outlook rule ..... استخدام Outlook rule	1.10
71	Zimbra Anti-spam system .....	2
74	الباب الثاني الدراسة النظري.....	
76	الفصل الرابع wordnet && wordnet domains .....	
76	wordnet3 ..... معجم wordnet3	1
81	مجالات Extended Wordnet Domains .....	2
82	الفصل الخامس RDF && DBpedia && Linked Data .....	
82	RDF ..... ثلاثيات RDF	1
84	DBpedia .....	2
85	Linked Data .....	3
86	الفصل السادس تحليل وفهم المحتوى.....	
87	استبدال الضمائر بالكلمات الأصلية التي تعود عليها هذه الضمائر Coreferencing .....	1
88	إعادة الكلمات إلى أصلها Lemmatization .....	2
88	إزالة غموض الكلمات Word Sense Disambiguation .....	3
90	خوارزمية Lesk التقليدية.....	3.1
92	التحسينات على خوارزمية Lesk .....	3.2
92	استخدام شبكة دلالية ..... 3.2.1	
95	ايجاد التقاءات بين تعريفين .....	3.2.2

96.....	3.2.3      استراتيجية عمل خوارزمية Global Disambiguation	
100.....	3.3      خوارزمية Hood	
102.....	3.3.1      فك غموض الكلمة وتصنيف المستندات	
102.....	3.3.2      إنشاء (Hood Construction) Hood	
104.....	3.3.3      فك الغموض	
104.....	3.4      إزالة الغموض بالاستعانة ب WordNet Tagged glosses	
104.....	3.4.1      لمحه عن مشروع WordNet Semantically Tagged glosses	
105.....	3.5      الخوارزمية المتبعة لإزالة الغموض	
107.....	3.5.1      ايجابيات و سلبيات هذه الخوارزمية	
107.....	4      مقارنة بين الخوارزميات الأربع المستخدمة في إزالة الغموض	
108.....	4.1      طريقة تحليل وفهم المحتوى اعتمادا على معالجة اللغات الطبيعية	
108.....	4.1.1      قراءة صفحة الإنترنت المطلوبة وتحويلها إلى ملف نصي	
108.....	4.1.2      تجزئة الكلمات	
108.....	4.1.3      إزالة المحارف الغربية	
108.....	4.1.4      تقسيم النص إلى جمل	
108.....	4.1.5      المعالجة الدلالية وعملية الفلترة	
110.....	4.2      طريقة فهم المحتوى اعتمادا على ثلاثيات RDF	
112.....	الباب الثالث مرحلة التحليل	

الفصل السابع وثيقة دراسة متطلبات النظام.....	114	
تعريف المشكلة المدروسة .....	114	1
الغاية من بناء هذا المنتج.....	115	2
تحديد الجهة المهمة بالمشروع.....	115	3
أجزاء المشروع الأساسية.....	115	4
الفصل الثامن حالات استخدام النظام .....	118	
مقدمة.....	118	1
مخطط حالات الاستخدام الأولى.....	118	2
<b>2.1 مخطط حالات استخدام مخدم البروكسي Smaz Proxy</b>	119	
<b>2.2 مخطط حالات استخدام مصنف الأخبار Smaz Filter</b>	120	
<b>3 حالات استخدام مخدم البروكسي Smaz Proxy</b>	121	
3.1 وضع إعدادات النظام.....	121	
3.2 تعديل إعدادات النظام.....	122	
3.3 إضافة قاعدة فلترة.....	123	
3.4 تعديل قاعدة فلترة .....	124	
3.5 طلب صفحة ويب.....	125	
3.6 حذف قاعدة فلترة .....	126	
3.7 عرض محتوى ملف القواعد .....	127	
3.8 عرض محتوى ملف الإعدادات .....	128	

129.....	<b>Smaz Filter</b>	4
129.....	<b>Smaz Reader</b>	5
129.....	عرض الأخبار التابعة لمجال معين .....	5.1
130.....	تحديد المجالات التي يرغب بعرض الأخبار التابعة لها	5.2
131.....	وضع إعدادات التطبيق.....	5.3
132.....	عرض المجالات المتواجدة .....	5.4
133.....	إضافة قناة أخبار.....	5.5
134.....	عرض القنوات المتواجدة .....	5.6
135.....	حذف قناة.....	5.7
136.....	وضع إعدادات النظام.....	5.8
الفصل التاسع مخطط الصنوف المفاهيمي ..... 140.....		
141.....	<b>Smaz Proxy</b>	1
142.....	توصيف الصنوف المفاهيمية لمخدم البروكسي .....	2
146.....	<b>Smaz Reader</b> للقارئ	3
147.....	مخطط الصنوف المفاهيمية للقارئ .....	4
الباب الرابع الدراسة التصميمية .....		
148.....	الفصل العاشر مخطط الصنوف .....	الباب الرابع الدراسة التصميمية .....
150.....	<b>Smaz Proxy</b>	1
150.....	مخطط صنوف مخدم البروكسي .....	1
151.....	<b>Smaz Proxy</b>	2
151.....	شرح بعض صنوف مخدم البروكسي .....	2

151.....	<b>hoodDisambiguator</b>	2.1
152.....	<b>Word</b>	2.2
152.....	<b>Combination</b>	2.3
152.....	<b>Instance</b>	2.4
153.....	<b>DbpediaHelper</b>	2.5
153.....	<b>ExDomainsHelper</b>	2.6
154.....	<b>FileReaderHelper</b>	2.7
154.....	<b>GeneralHelper</b>	2.8
155.....	<b>StanfordCoreNLPHelper</b>	2.9
155.....	<b>DocDomainsFinder</b>	2.10
156.....	<b>SynsetDomain</b>	2.11
156.....	<b>JWIHelper</b>	2.12
156.....	<b>SemanticAnalyzer</b>	2.13
156.....	<b>SemanticAnalyzerImpl</b>	2.14
157.....	<b>Smaz Reader</b>	3
158.....	<b>Smaz Reader</b>	4
158.....	<b>RssChannel</b>	4.1
158.....	<b>RssItem</b>	4.2
159.....	<b>Domain</b>	4.3
159.....	<b>DataBaseHelper</b>	4.4
160.....	<b>SemanticAnalyzerDomain</b>	4.5
160.....	<b>WebHelper</b>	4.6
161.....	<b>Config</b>	4.7

الفصل الحادي عشر تصميم قاعدة المعطيات.....	164	
جداول توصيف الكيانات .....	164	1
العلاقات بين الكيانات .....	166	2
مخيط قاعدة البيانات .....	167	3
الباب الخامس مرحلة التحقيق .....	168	
الفصل الثاني عشر تحقيق النظام .....	170	
خدمة الويب Smaz web service .....	172	1
فكرة RMI .....	172	1.1
RMI Server .....	174	1.1.1
RMI Client .....	174	1.1.2
بنية ملف الإعدادات Config File .....	175	2
بنية ملف القواعد Rules File .....	178	3
خدم البروكسي Smaz Proxy .....	180	4
Icap server .....	180	4.1
مخيط العمل وأآلية الاتصال بين مكونات النظام .....	181	4.2
إضافة المتصفح Smaz Filter .....	182	5
القارئ Smaz Reader .....	183	6
اختبار أداء Smaz Proxy .....	183	
البيئة البرمجية .....	185	7

185.....	لغة البرمجة المستخدمة	7.1
187.....	jsp web service	7.2
187.....	ICAP	7.3
187.....	المكتبات المستخدمة	7.4
187.....	Apache Jena	7.4.1
188.....	Apache PDFBox	7.4.2
188.....	Apache POI	7.4.3
188.....	Boilerpipe	7.4.4
189.....	Jsoup	7.4.5
189.....	JWI	7.4.6
191.....	Stanford CoreNLP	7.4.7
192.....	SpotLight	7.5
194.....	الفصل الثالث عشر واجهات النظام	
195.....	Smaz Proxy	1
195.....	مخدم icap والبرمجة بداخله	1.1
196 .....	التأكد من أن المخدم متصل وقدر على معالجة الطلبات	1.2
197.....	طلب صفحة ومعالجتها وحجبها	1.3
198.....	وضع إعدادات المخدم	
199.....	إضافة المتصفح Smaz Filtering	2
199 .....	واجهة الإغاثة التي يتم من خلالها وضع الإعدادات	2.1

200.....	معلومات عن الإضافة.....	2.2
200.....	إضافة Smaz Filter إلى قائمة إضافات المتصفح.....	2.3
201.....	تطبيق الموبايل Smaz Reader.....	3
201 .....	واجهة عرض المجالات المختارة مسبقا	3.1
202.....	واجهة عرض الأخبار الخاصة بمجال معين	3.2
203.....	واجهة إعدادات التطبيق.....	3.3
204.....	واجهة عرض قنوات الأخبار.....	3.4
205.....	واجهة تعديل قناة أخبار.....	3.5
206.....	واجهة توضح اختيار المجالات المطلوب عرضها	3.6
207.....	واجهة توضع معلومات عن التطبيق	3.7
208.....	محرر القواعد Rule editor.....	4
208.....	واجهة إضافة قاعدة.....	4.1
209.....	واجهة اختيار مجال من المجالات وفق الهرمية الموضوعة	4.2
210.....	واجهة عرض القواعد الموجودة والتعديل عليها	4.3
212.....	الباب السادس مرحلة الاختبارات.....	
214.....	مقدمة.....	1
214.....	الاختبارات عبر مراحل تطوير النظام.....	1
214.....	اختبار الواحدة البرمجية.....	1.1
214.....	اختبار التكامل.....	1.1.1

215.....	اختبار النظم	1.1.2
215.....	اختبارات الجودة	1.1.3
218.....	الباب السابع الآفاق المستقبلية، الملحقات والمراجع .....	
220.....	الفصل الرابع عشر الآفاق المستقبلية .....	
222.....	الفصل الخامس عشر الملحق أـ مفرد المصطلحات .....	
226.....	الفصل السادس عشر الملحق بـ - فهرس المخططات.....	a
230.....	الفصل السابع عشر الملحق جـ - فيروس الجداول .....	
232.....	الفصل الثامن عشر الملحق دـ - المراجع .....	



## تمهيد

تشهد الشبكة العنكبوتية اليوم تطويراً ملحوظاً وتزايداً في أعداد الواقع التي تنضم يومياً إلى لائحة الواقع على شبكة الإنترنت، والتي تتتنوع في محتواها فمنها الترفيهي ومنها الثقافي ومنها التعليمي فضلاً عن شبكات التواصل الاجتماعي وغير ذلك من المجالات، هذا النمو المتزايد جعل من الصعب السيطرة على محتوى صفحات الإنترنت وضمان أن ما يتم تصفحه هو ضمن مجال معين نرغب به.

من هنا أتت الحاجة إلى فلترة صفحات الإنترنت، فمثلاً لضمان عدم القدرة على الوصول إلى الواقع الإباحية يتم حظر قائمة من الواقع الإباحية وبالتالي عند طلب المستخدم لهذا الموقع لن يتم عرض تلك الصفحة من قبل المتصفح، لكن هذه الطريقة في الحجب غير مجديّة نظراً لعدم قدرتنا على حصر الواقع التي تدرج تحت مجال معين، فضلاً عن ظهور موقع جديدة يومياً، أو أن يتم الحجب من خلال تحديد قائمة من الكلمات الغير مرغوب بها وبالتالي عند طلب صفحة تحوي كلمة من هذه الكلمات لن يتم عرض محتواها، لكن هذه الطريقة ليست عملية أيضاً ففي حالات قد ترد كلمات مرادفة للكلمات الغير مرغوب بها ومع ذلك سيعتمد عرض الصفحة لعدم ورود نفس الكلمات التي تم تحديدها سابقاً، وبالتالي نحن بحاجة إلى طريقة فلترة آنية تتم لحظة طلب الموقع لتقام عليه عمليات معالجة وبناءً على تلك المعالجة الدلالية يتم عرض الصحة أم لا.

نقدم في هذا المشروع توصيفاً لطريقة عملية آنية تعتمد على الفهم الدقيق لمحتوى صفحة الإنترنت لحظة طلبها من قبل المستخدم حيث نقوم بالحصول على النص الموجود ضمن صفحة الإنترنت وفهمه من خلال عمليات معالجة نحوية دلالية توصلنا في النهاية إلى فهم محتوى الصفحة ومعرفة المجالات التي تدور حولها وفق نسب مئوية تعبر عن مدى انتمام الصفحة إلى كل مجال وبناءً على تلك النسبة يتم أخذ قرار عرض تلك الصفحة للمستخدم أم لا وفق قواعد مسبقة موضوعة من قبل خبراء وبما يتماشى مع حاجات الشركة المستخدمة لهذا النظام.



# الفصل الأول

## مدخل إلى المشروع

1 مقدمة

بدأت المكتبات باستخدام فلاتر الإنترنت في أواخر 1990 بسبب ضغوط المجتمعات وقانون حماية الإنترنت للطفلة

Children's Internet Protection Act (CIPA) وهو القانون الاتحادي الذي يتطلب من جميع أجهزة الكمبيوتر

في المكتبة العامة أن تحوي فلاتر لصفحات الإنترنت التي سيتم طلبها من قبل مستخدمي هذه المكان، حيث تقول إحدى الدراسات

الإحصائية أن نسبة استخدام الفلاتر في المكتبات قد ازدادت من 25٪ في عام 2000 إلى 65٪ في عام 2005.

في عام 1996 ، أصدر الكونغرس الأميركي قانون حظر الواقع الإباحية والواقع ذات المحتوى السيء من الإنترنت ، وفي

عام 1997 وقفت جماعة الحريات المدينة ضد ذلك القانون وألغي بعد ذلك ، وهكذا أصبح من الضروري تصفية المحتوى.

منذ 1990 s ، وقفت العديد من المجموعات ضد استخدام الشركات لتصفية المحتوى ، نددت بالاستخدام المفرط لفلاتر

الإنترنت وبالتالي حجب موقع ليس هناك أسباب واضحة وراء حجبها ، ولكن هناك حاجة ضرورية لفلترة المحتوى من قبل الآباء

لحماية أطفالهم من الواقع السيئة ، أصبحت شركات تجارية كبرى تعمل على فلترة المحتوى ، وقد شهد السوق نموا مضطربا ضمن

هذا المجال ، وحتى نمو 20 في المئة خلال اقتصاد الركود في عام 2009.

تطورت خلال هذه الفترة طرق الفلترة المستخدمة بداء من الحجب اعتماداً على قائمة سوداء تضم مجموعة من المواقع التي لا نرغب بعرضها، ومن ثم الحجب من خلال مجموعة من الكلمات لمجرد ورود إحدى هذه الكلمات ضمن صفحة الإنترنت المطلوبة لن يتم عرض الصفحة، والحجب اعتماداً على meta tags محددة لمجرد ورود هذه الكلمات المفتاحية ضمن الكلمات المفتاحية الخاصة بالصفحة يتم الحجب، والحجب اعتماداً على تعابير نظامية وغير ذلك من الطرق المستخدمة في الفلترة.

تعاني هذه الطرق من عدم جديتها ودققتها فلن نستطيع بهذه الطرق الحجب الفعلي لمحظى كامل غير مرغوب به وبالتالي لابد من طرق أكثر عملية وفعالية بنفس الوقت، هنا لابد من الفهم الدلالي الدقيق لمحظى الصفحة واتخاذ قرار الحجب لحظة طلب الصفحة بناءً على محتواها.

## 2 عرض المشكلة

بعد التزايد الكبير لموقع الإنترت وعدم القدرة على إحصائها، أصبح من غير الممكن متابعة كل موقع الإنترت ومعرفة محتواها ومضمونها، وبالتالي أصبح من الصعب القيام بعمليات مسبقة لمعالجة المحتوى من فلترة وتصنيف وغير ذلك من العمليات.

فلترة صفحات الإنترت كانت عبارة عن حصر المواقع التي تزيد حجبها ووضعها ضمن قائمة سوداء وعند طلب صفحة موجودة ضمن القائمة يتم حجبها عن المستخدم، لكن هذه الطريقة لم تعد مجديّة كون الشبكة العنكبوتية في تطور ونمو مستمر لا يمكن السيطرة عليه.

فلترة صفحات الإنترت من خلال قائمة من الكلمات الغير مرغوب بها غير مضمونة أيضاً، فيمكن ورود كلمات ضمن صفحة الإنترنت المطلوبة بمعنى الكلمة الغير مرغوب بها ويتم عرض الصفحة لعدم ذكر الكلمة صراحة.  
إذا كل عمليات معالجة المحتوى من فلترة وتصنيف وغير ذلك تتطلب فهم للمحتوى، عملية الفهم المسبق لكل المحتوى عملية ليست منطقية ولا يمكن تطبيقها، واللجوء إلى عمليات المعالجة اليدوية غير مجدي.

### 3 أهداف المشروع

يهدف المشروع إلى تقديم خدمة تعمل على فهم المحتوى بشكل آلي آني، تقوم هذه الخدمة باستخراج النص من الصفحة المطلوبة، معالجة هذا النص من خلال إجراء عمليات معالجة لغات طبيعية متتالية، ومن ثم الحصول على معاني كلمات من خلال الاستعانة بالمعالج والأنطولوجيات المشهورة، ومن ثم الحصول على المجالات التي تنتهي لها الصفحة، أي معرفة أهم المواضيع والمجالات التي تتحدث عنها الصفحة المطلوبة، بهذه الطريقة بتنا قادرين على الفهم الدلالي لمحتوى الصفحة ومعرفة مضمونها، هذا الفهم الدلالي يتم بشكل آلي وبالتالي تم التخلص من طرق المعالجة اليدوية الغير مجدية، بهذه الطريقة بتنا قادرين على القيام بعمليات على المحتوى من فلترة وتصنيف وغير ذلك.

يقدم المشروع إذاً طريقة دلالية آنية لفهم المحتوى، ومن ثم تم الاستفادة من هذه الطريقة في فلترة المحتوى من خلال تطبيقين chrome extention و proxy server ، وتصنيف المحتوى من خلال تطبيق موبайл يتم فيه تصنیف الأخبار ضمن مجموعة من المجالات كما سنرى لاحقا.

## 4 إدارة المشروع

نعرض في هذا الفصل الموارد المتنوعة البشرية والعتادية التي احتاجها المشروع، ثم سنقوم بعرض المهام الرئيسية في المشروع، إضافة إلى تحديد الفترة اللازمة لتحقيق كل مهمة وذلك من خلال جدول يوضح الفترة الزمنية المحددة لكل مهمة.

### 4.1 الموارد المتاحة

#### 4.1.1 الموارد العتادية

أربعة حواسيب محمولة مجهزة بنظام Microsoft windows 7 وحواسيب منهما مجهزان بنظام التشغيل Linux.

#### 4.1.2 الموارد البرمجية

- بيئة netbeans java لتطوير البرمجيات؛

- بيئة Eclipse،

Java for android -

- بيئة java script،

- برنامج Enterprise Architect لتصميم النظام هندسيا.

#### **4.1.3 الموارد البشرية**

##### **• فريق عمل المشروع**

تضمن فريق العمل أربع أشخاص:

- أسامة سليق؛

- آنا عجميان؛

- بلال الهلال الشريفي؛

- لما موازني.

##### **• الإشراف العلمي**

تضمن فريق الإشراف العلمي:

- الدكتور خليل عجمي؛

- الدكتور سامي خيمي؛

- المهندس محمد الساطي.

## **4.2 المهام الرئيسية**

#### **4.2.1 مرحلة الدراسة النظرية**

- دراسة الخوارزميات المستخدمة في إزالة غموض الكلمات؛

- دراسة حول طرق فهم النص دلاليًا؛

- دراسة عن أشهر البرامج المستخدمة في فلترة صفحات الإنترنت وطرق عملها.

مرحلة التحليل 4.2.2

- وضع حدود المشروع؛
- تحديد متطلبات المشروع؛
- تحديد حالات الاستخدام.

مرحلة التصميم 4.2.3

- تصميم قاعدة معطيات النظام؛
- تصميم الصنوف الالزامية لتحقيق النظام.

مرحلة التحقيق 4.2.4

- تحقيق قاعدة المعطيات؛
- تحقيق صنوف النظام؛
- تحقيق الرماد المدرسي.

مرحلة الاختبارات 4.2.5

- اختبار الوحدات البرمجية؛
- اختبار التكامل؛
- اختبار النظام؛
- اختبار الجودة .

الجدول الزمني للمهام 4.2.6

نعرض فيما يلي جدول زمني يوضح المدة الزمنية لكل مرحلة من مراحل المشروع حيث تتضمن كل مرحلة التقرير والتوثيق المرافق للعمل.

الرقم	المهمة	المدة	تاريخ البدء	تاريخ الانتهاء
1	دراسة خوارزميات إزالة غموض الكلمات.	15	1/3/2013	15/3/2013
2	دراسة حول فهم النص دلالي.	20	16/3/2013	5/4/2013
3	دراسة عن برامج فاترة الإنترنت وطرق عملها.	10	6/4/2013	15/4/2013
4	وضع حدود المشروع.	5	16/4/2013	20/4/2013
5	تحديد متطلبات المشروع.	10	21/4/2013	30/4/2013
6	تحديد حالات الاستخدام.	5	1/5/2013	5/5/2013
7	تصميم قاعدة معطيات النظام.	5	6/5/2013	10/5/2013
8	تصميم الصفوف اللازمة لتحقيق النظام.	10	11/5/2013	20/5/2013
9	تحقيق قاعدة المعطيات.	5	21/5/2013	25/5/2013
10	تحقيق صفوف النظام.	5	26/5/2013	30/5/2013
	تحقيق الرماز المصري	20	1/6/2013	20/6/2013

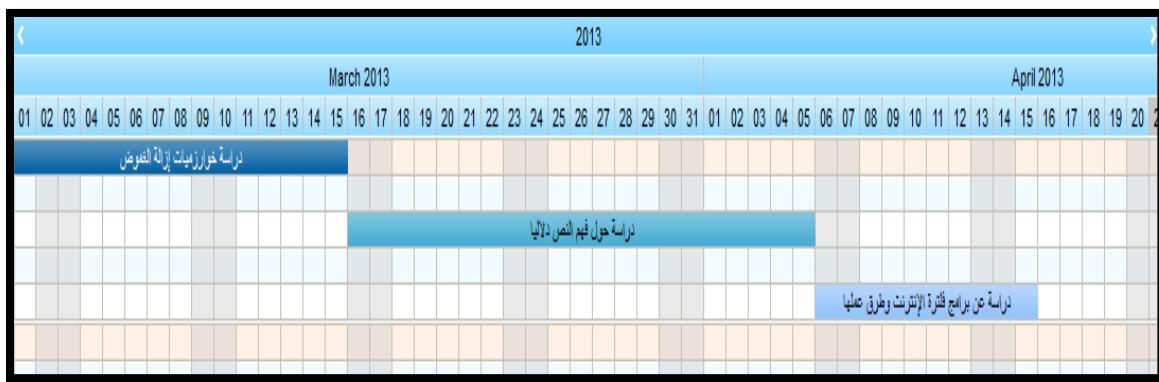
30/6/2013	21/6/2013	10	مرحلة الاختبارات	11
10/7/2013	1/7/2013	10	العرض التقديمي	12

**جدول 1 توزيع المهام خلال فترات زمنية محددة**

## 4.3 مخططات غانت

تظهر هذه المخططات توزع المهام زمنياً والعلاقات المتبادلة بين هذه المهام.

### 4.3.1 مخطط غانت لمرحلة الدراسة النظرية



الشكل 1 مخطط غانت لمرحلة الدراسة النظرية

### 4.3.2 مخطط غانت لمرحلة التحليل



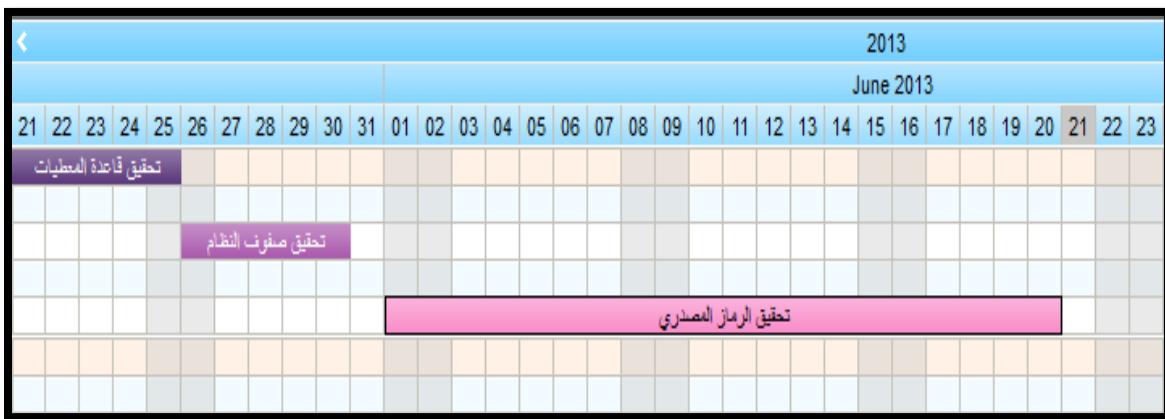
الشكل 2 مخطط غانت لمرحلة التحليل

مخطط غانت لمرحلة التصميم 4.3.3



الشكل 3 مخطط غانت لمرحلة التصميم

مخطط غانت لمرحلة التحقيق 4.3.4



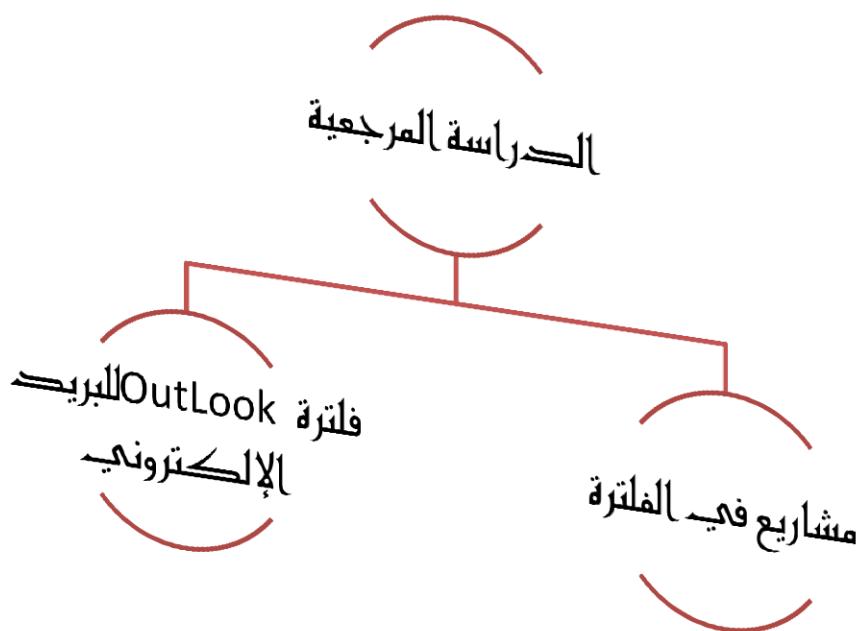
الشكل 4 مخطط غانت لمرحلة التحقيق



الشكل ٥ مخطط غانت لمرحلة الاختبار

# الباب الأول

## الدراسة المرجعية



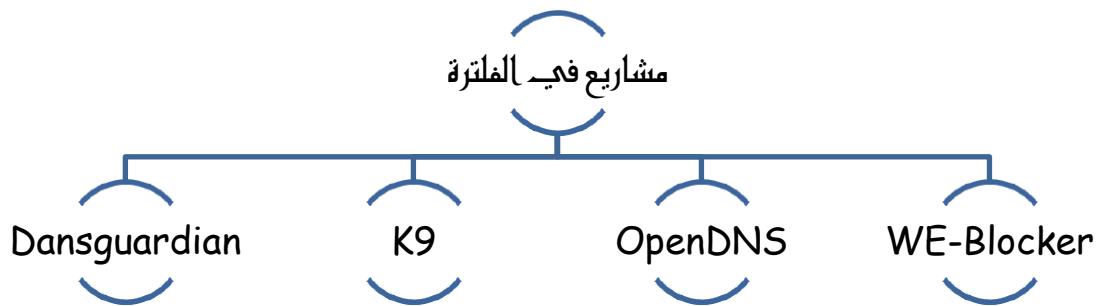
الشكل 6 النقاط الأساسية في الدراسة المرجعية



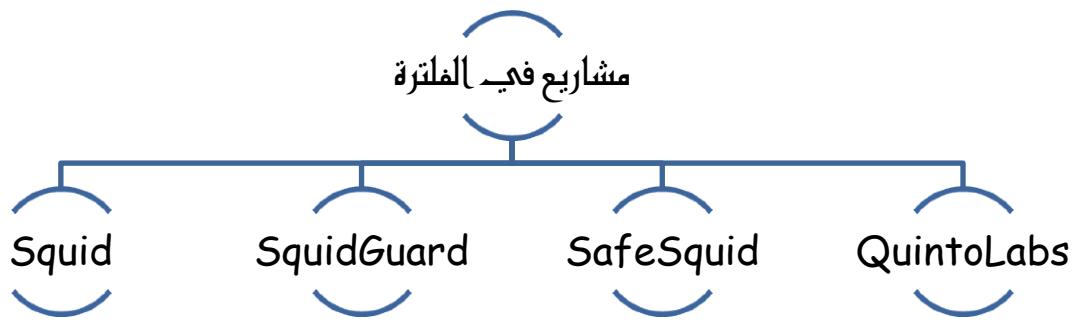
## الفصل الثاني

### مشاريع فيه فلترة سفادات الإنترن特

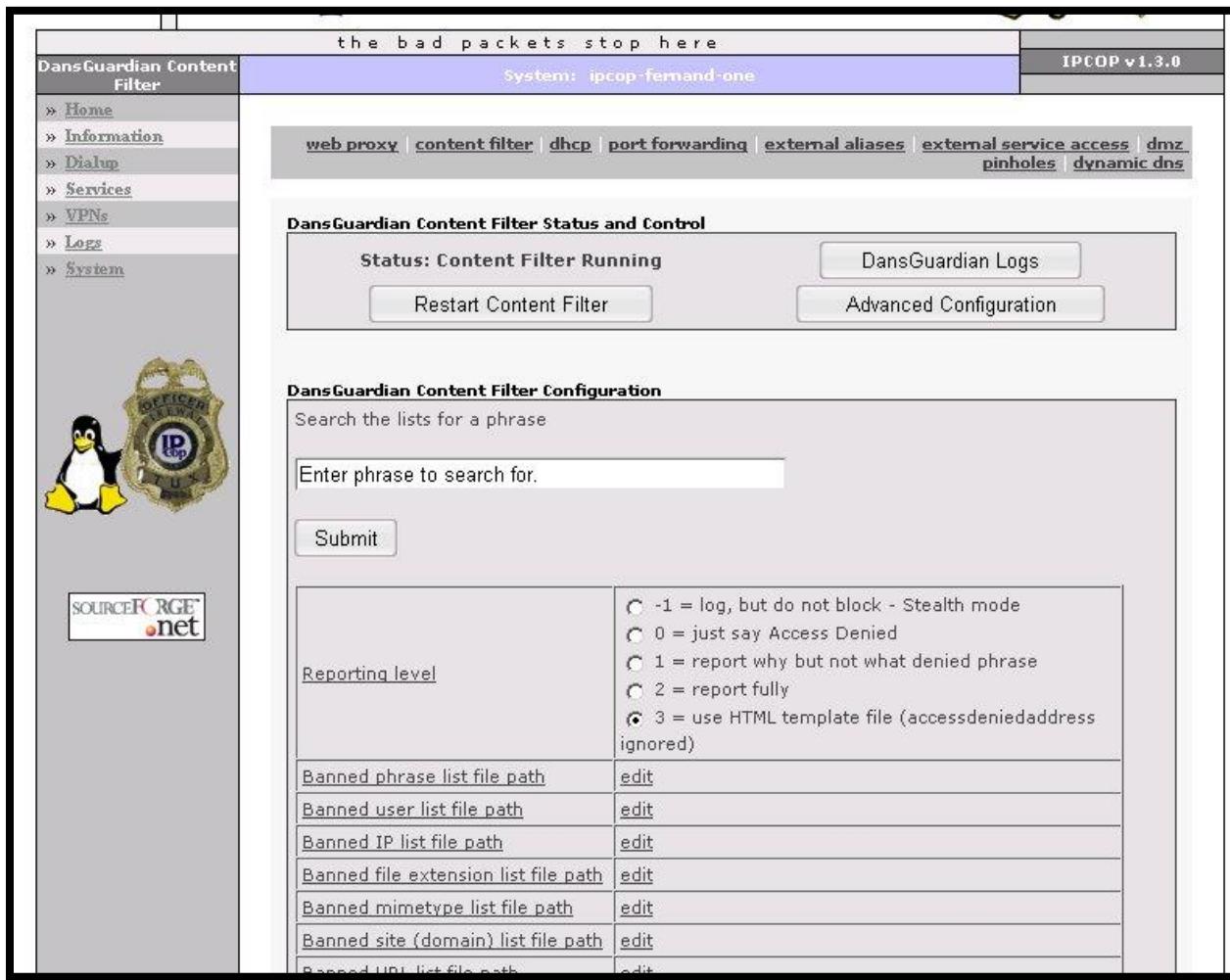
يتناول المشروع مجموعة من التطبيقات الناتجة عن فهم وتحليل للمحتوى تم استخدامه في فلترة صفحات الإنترن特 من خلال تطبيقيين اثنين، كما تم استخدامه في تصنيف المحتوى من خلال تطبيق موبайл، نعرض في هذا الفصل مجموعة من البرامج والتطبيقات التقليدية المستخدمة في فلترة صفحات الإنترن特، لنرى كيف تتم الفلترة في الواقع وفي معظم الأحيان نجد أن الطرق المستخدمة هي طرق تقليدية، وأن الخوارزميات المتبعة هي خوارزميات بسيطة لا تفي بالغرض المرجو.



الشكل 7 النقاط الأساسية في مشاريع الفلترة



(Cross Platform, Free) DansGuardian 1.1



الشكل 9 صورة توضح واجهة DansGuardian

- مجاني (في حال الاستخدام غير التجاري) ومتاح المصدر؛
- يدعم أنظمة التشغيل التالية: Linux, FreeBSD, OpenBSD, NetBSD, Mac OS X, HP-UX, Solaris

- يعمل كإضافة على بعض خدمات البروكسي مثل Squid ،

- يقوم بالفلترة من خلال :

1- فلترة الصفحات ليس فقط بناءً على قوائم عناوين معدة مسبقاً كما تفعل باقي الفلاتر، وإنما أيضاً يقوم بالفلترة بناءً

على محتوى الصفحات؛

2- يتيح إمكانية الفلترة بناءً على العناوين URL filtering أو باستخدام black/white lists

و هنا يتم التمييز بين نوعين للفلترة: regular expressions

a. فلترة موقع معين بشكل كامل؛

b. فلترة صفحات معينة من موقع ما وهذه الطريقة في الفلترة غير مجدية كثيراً بسبب التزايد المستمر والسرع

موقع الإنترنـت و صعوبة تعديل قوائم الحجب .black/white lists

3- يتيح إمكانية فلترة المحتوى باستخدام regular expressions ،

4- يتيح إمكانية فلترة المحتوى باستخدام ما يسمى DansGuardian Phraselists ، حيث أن

الافتراضية يأتي ضمناً بأعداد ضخمة من هذه العبارات مقسمة إلى أصناف رئيسية، وبالطبع يمكن تعديليها، ولهذه

العبارات نوعين:

Banned Phrases .a

في حال ورود أي عبارة من هذه العبارات في الصفحة يتم حجب الصفحة بشكل كامل دونأخذ أي شيء آخر

بعين الاعتبار، لذلك يجب توخي الحذر عند التعديل على العبارات من هذا النوع.

مثلاً في حال ورود الكلمة sex يتم حجب الصفحة بشكل كامل، وتتم صياغة هذا النوع على الشكل:

<sex>

Weighted Phrases .b

هنا يتم مقابله كل عبارة بمجموع نقط، ويتم في النهاية حساب مجموع النقط التراكمي للصفحة و في حال تجاوزت حد معين (يمكن تعديله) يتم حجب الصفحة.

مثلاً إذا وردت الكلمة education مع الكلمة sex هذا سيخفف من الأثر السلبي وبالتالي يجب إضافة عدد سالب إلى المجموع التراكمي للصفحة و يتم صياغة العبارة من هذا النوع على الشكل التالي :

<sex><50>

<sex>,<education><-30>

أي يتم طرح 30 من المجموع التراكمي للصفحة في حال وردت كلا الكلمتين السابقتين في الصفحة وهذه الطريقة وضوحاً تعتبر أكثر دقة من سابقتها .

5- يتيح إمكانية تعديل ال URL أو محتوى الصفحات في بعض الأحيان، مثلاً في حال البحث في Google يمكن أن يتم تعديل ال URL لاجبار Google Safe Search على العمل حتى لو لم يقوم الشخص الذي

طلب الصفحة بتفعيل هذه الميزة صراحة؛

6- يتيح إمكانية تفعيل ميزة POST limiting والتي تسمح بالتحكم برفع الملفات أو حجبها بشكل كامل؛

PICs web-site rating filtering -7

أي يمكن أن تتم الفلترة على مستوى الصور؛

Anti-virus filtering -8

يتم اختبار الملفات قبل تحميلها فيما إذا كانت خبيثة أم لا؛

Meta tags filtering -9

تمت الفلترة بناء على ال meta tags الخاصة بال html ،

File extension and file type (MIME) filtering -10

أي يتم حجب تحميل الملفات و الصفحات التي لها أنماط محددة يتم تحديدها مسبقاً [1]

The screenshot shows the K9 Web Protection Administration interface. At the top, there's a navigation bar with icons for Home, View Internet Activity, Setup, and Get Help. On the left, a sidebar lists various settings: Web Categories to Block, Time Restrictions, Web Site Exceptions, Blocking Effects, URL Keywords, Other Settings, and Password/Email. The main content area is titled "Web Categories to Block". It includes a note to set categories to block or allow, with a "More Help..." link. There are four radio button options: "High" (protects against default-level categories), "Default" (protects against adult content, security threats, illegal activity, sexually-related sites, and online community sites), "Moderate" (protects against adult content, security threats, and illegal activity), "Minimal" (protects against pornography and security threats), "Monitor" (allows all categories - only logs traffic), and "Custom" (selects own categories). A table lists currently blocked categories in three columns: Abortion, Adult / Mature Content, Alternative Sexuality / Lifestyles; Alternative Spirituality / Occult, Extreme, Gambling; Hacking, Illegal / Questionable, Illegal Drugs; Intimate Apparel / Swimsuit, LGBT, Nudity; Open Image / Media Search, Peer-to-Peer (P2P), Personals / Dating; Phishing, Pornography, Proxy Avoidance; Sex Education, Social Networking, Spyware / Malware Sources; Spyware Effects, Suspicious, Violence / Hate / Racism. Below the table, there are two boxes: "Time Restrictions" (set times to block or allow Web access) and "View Activity Summary" (display an overview of blocked Web sites and other Internet events on your computer).

الشكل 10 صورة توضح واجهة K9

- مجاني من أجل الاستخدام الفردي والمنزلي، غير مجاني من أجل الشركات؛
- لا يعمل مع نظام التشغيل Linux؛
- يدعم أنظمة التشغيل Windows, Mac و Android كمتضيّع؛

- يقوم K9 بالفلترة من خلال:

Web Site Exceptions Blocks -1 منع أو السماح بموقع معينة تختارها؛

URL Keywords Blocks -2 تحديد كلمات معينة وعند محاولة الدخول على اسم موقع يحتوى هذه الكلمة سيتم منع الدخول تلقائياً، فعلى سبيل المثال لنفرض حذفنا كلمة "casino" فعند ورود URL حاوي

على هذه الكلمة مثل bestCasino.com فيتم حذفه مباشرة كونه يحتوى على كلمة محجوبة؛

Time Restrictions Blocks -3 تحديد أوقات الدخول إلى الإنترنت؛

Timeout Blocks -4 تحديد عدد المواقع الغير مسموح الدخول إليها خلال فترة زمنية محددة في حال

تجاوز ذلك العدد يتم مثلا حجب الدخول إلى أي موقع لمدة نصف ساعة على سبيل المثال؛

Web Categories to Block -5 تحديد مستوى الفلترة العام هناك خيارات:

yo tube والـ الشـات شيء كل تقريباً يمنع High

Default: الإفتراضي، يمنع المحتوى السيء، فمثلا يقوم بحجب المواقع بناءً على تصنيف مسبق بحيث

يقوم بحذف موقع playboy.com كونه مصنف على أنه من النوع Pornography.

كما يمكنه منع حجب المواقع المصنفة على أنها Sex Educations مثلا من خلال تعديل بسيط في

الإعدادات كالسماح بفتح كل المواقع ذات المحتوى التعليمي والشرط أن يملك الشخص الذي يقوم بالتعديل

صلاحيات الإدارة؛

6- يمكن إعداد فاتر خاص بالمستخدم عن طريق Custom فمثلا يقوم المستخدم بتحديد التصنيفات التي يرغب

بحجبها والمواقع التي يريد والكلمات التي يريد حجبها من عنوان المواقع؛

7- عند طلب URL موقع ما غير محظوظ صراحة ولا يحوي على كلمات محظوظة:

✓ يقوم K9 بالبحث عن الموقع المطلوب ضمن الذاكرة الخاصة به cache memory ، في حال حدوث

مطابقة لا يقوم بسؤال BCWF عن تصنيفه ؛

✓ وإنما يقوم بطلب استعلام من BCWF لتقوم بدورها بالبحث عنها ضمن قاعدة المعطيات الخاصة بها ، في

حال حدوث مطابقة تقوم بإرسال التصنيف إلى K9 ، وإنما سيقوم BCWF بجلب محتوى الصفحة

وعمل تقييم آني لها بناء على DRTR ، حيث يقوم بعملية تصنيف مباشرة على الموقع عن طريق

خوارزميات تحليلية سريعة اعتماداً على خوارزمية ذكية عالية المستوى هي عملية التعليم Machine Learning

، حيث يعتمد على الكم الهائل من الواقع المتوفرة لديه كعينات تدريب .

وهذه العملية لا تأخذ وقت أكثر من 70 ثانية ، وهي تعمل بشكل ممتاز على الواقع التي تكون من النوع

Pornography حيث أن هذا النوع من الواقع هو الأكثر تقييماً ضمن قاعدة البيانات الخاصة بهذا

النظام ، وأيضاً فهو يعتمد على links الموجودة ضمن الصفحات بحيث يبحث عن هذه الترابطات ضمن

قاعدة بيانياته الخاصة ، وفي حال وجد أن هذه الترابطات تشير إلى أحد الواقع المخزنة لديه يقوم بمقارنته

معه مباشرة ويعطيه نفس التصنيف . [2]

### Dynamic Real-Time Rating™ (DRTR)

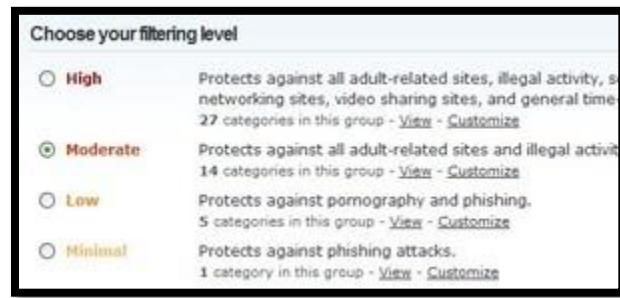
✓ التقييم динамический في الوقت الحقيقي ؛

✓ هي طريقة قوية ومتينة وأقوى من طرق الفلترة التقليدية المعتمدة على الكلمات المفتاحية ؛

✓ تعتمد على خوارزمية تحليل إحصائي وذكاء اصطناعي لتقييم صفحات الإنترنت الجديدة التي لم يتم تقييمها مسبقاً ؛

✓ تستطيع ال DRTR تحديد تصنيف صفحة الإنترنت من دون تدخل بشري ؛

## (Cross Platform, Free) OpenDNS 1.3



الشكل 11 صورة توضيح واجهة OpenDNS

- خدمة توفر DNS Server مجاني ولكن أيضاً بميزات تمنحها أغلب موفري الخدمات بمقابل مادي وهي "فلترة المحتوى" ؟
- تقوم بكتابة ips الخاصة بال DNS Server سواء كان حاسوب أو راوتر، وتقوم بالتسجيل في الموقع وتدوين ip خاص بالشبكة وتعطي تلك الشبكة اسم، وبعد ذلك تحدد ما هو المرغوب به وما هو غير المرغوب، وفي حال حاول أحد الدخول على أحد الأشياء التي قمت بمنعها يتم توجيهه إلى صفحة تخبره برسالة معينة دون تحميل الصفحة المطلوبة ؛
- القدرة على حظر الواقع عن كل المتصلين بالشبكة سواء عن طريق حاسب أو هاتف أو أي جهاز لديه القدرة على الإتصال بالإنترنت ؛
- يمكن إرغام مستخدمي الشبكة عن كتابة عناوين معينة بأن توجههم إلى ips معينة مثلاً إذا كتب شخص ما kernel.org يتم توجيهه إلى microsoft.con
- إمكانية حجب موقع تحتوي كلمات معينة وتجاهل الآخر؛ كما أن هناك خاصية الواقع المسموحة والمنوعة ؛
- الحماية من هجمات السييرفاتر والإعلانات الضارة ومن بعض التغرات في الواقع ؛

- القدرة على مراقبة الشبكة الخاصة ومعرفة أكثر المواقع تصفحاً وأكثرها حظراً وأكثر الأوقات ازدحاماً وما إلى ذلك؛
- توفر بعض التسهيلات في عملية التصفح مثل البحث التلقائي وتصحيح بعض الأخطاء في العناوين وإمكانية وضع الاختصارات؛
- هذا بالإضافة إلى أنها تقوم بوظيفة الـ DNS الإعتيادية مع العلم أن الـ DNS سريع ومستقر؛
- إمكانية تحديد تصنيفات ليتم حظرها كما هو موضح في الشكل:

**Web Content Filtering**

**Choose your filtering level**

- High** Protects against all adult-related sites, illegal activity, social networking sites, video sharing sites, and general time-wasters. 26 categories in this group - [View](#) - [Customize](#)
- Moderate** Protects against all adult-related sites and illegal activity. 13 categories in this group - [View](#) - [Customize](#)
- Low** Protects against pornography. 4 categories in this group - [View](#) - [Customize](#)
- None** Nothing blocked.
- Custom** Choose the categories you want to block.

<input type="checkbox"/> Academic Fraud	<input type="checkbox"/> Adult Themes	<input checked="" type="checkbox"/> Adware
<input type="checkbox"/> Alcohol	<input type="checkbox"/> Auctions	<input type="checkbox"/> Automotive
<input type="checkbox"/> Blogs	<input type="checkbox"/> Business Services	<input type="checkbox"/> Chat
<input type="checkbox"/> Classifieds	<input type="checkbox"/> Dating	<input type="checkbox"/> Drugs
<input type="checkbox"/> Ecommerce/Shopping	<input type="checkbox"/> Educational Institutions	<input type="checkbox"/> File storage
<input type="checkbox"/> Financial institutions	<input type="checkbox"/> Forums/Message boards	<input type="checkbox"/> Gambling
<input type="checkbox"/> Games	<input type="checkbox"/> Government	<input type="checkbox"/> Hate/Discrimination
<input type="checkbox"/> Health	<input type="checkbox"/> Humor	<input type="checkbox"/> Instant messaging
<input type="checkbox"/> Jobs/Employment	<input checked="" type="checkbox"/> Lingerie/Bikini	<input type="checkbox"/> Movies
<input type="checkbox"/> Music	<input type="checkbox"/> News/Media	<input type="checkbox"/> Non-profits
<input checked="" type="checkbox"/> Nudity	<input type="checkbox"/> P2P/File sharing	<input type="checkbox"/> Parked Domains
<input type="checkbox"/> Photo sharing	<input type="checkbox"/> Podcasts	<input type="checkbox"/> Politics
<input checked="" type="checkbox"/> Pornography	<input type="checkbox"/> Portals	<input checked="" type="checkbox"/> Proxy/Anonymizer
<input type="checkbox"/> Radio	<input type="checkbox"/> Religious	<input type="checkbox"/> Research/Reference
<input type="checkbox"/> Search engines	<input checked="" type="checkbox"/> Sexuality	<input type="checkbox"/> Social networking
<input type="checkbox"/> Software/Technology	<input type="checkbox"/> Sports	<input type="checkbox"/> Tasteful
<input type="checkbox"/> Television	<input type="checkbox"/> Tobacco	<input type="checkbox"/> Travel
<input type="checkbox"/> Video sharing	<input type="checkbox"/> Visual search engines	<input type="checkbox"/> Weapons
<input type="checkbox"/> Webmail		

Looking for security categories?

الشكل 12 صورة توضح التعامل مع openDNS

- في حال كان هناك بعض المواقع الاستثنائية التي تريد منها ولا تريد منع تصنيفاتها كاملة فيمكن عمل ذلك، كما يمكن

إلغاء حظر بعض المواقع التي تم حظر تصنيفاتها كما في الشكل:

- إمكانية تحديد شكل رسالة الحظر من حيث النص الذي سيظهر والصورة المعروضة؛

- قاعدة البيانات الخاصة بالخدمة لم تنشأ إلا عن طريق إبلاغ المستخدمين عن الموقف، وعند وصول من بلغوا عن الموقع

بتقنيات معينة إلى عدد معين يتم قبول هذا التصنيف؛

- مجاني ومفتوح المصدر.<sup>[3]</sup>

#### (Windows XP or earlier, Free) We-Blocker 1.4

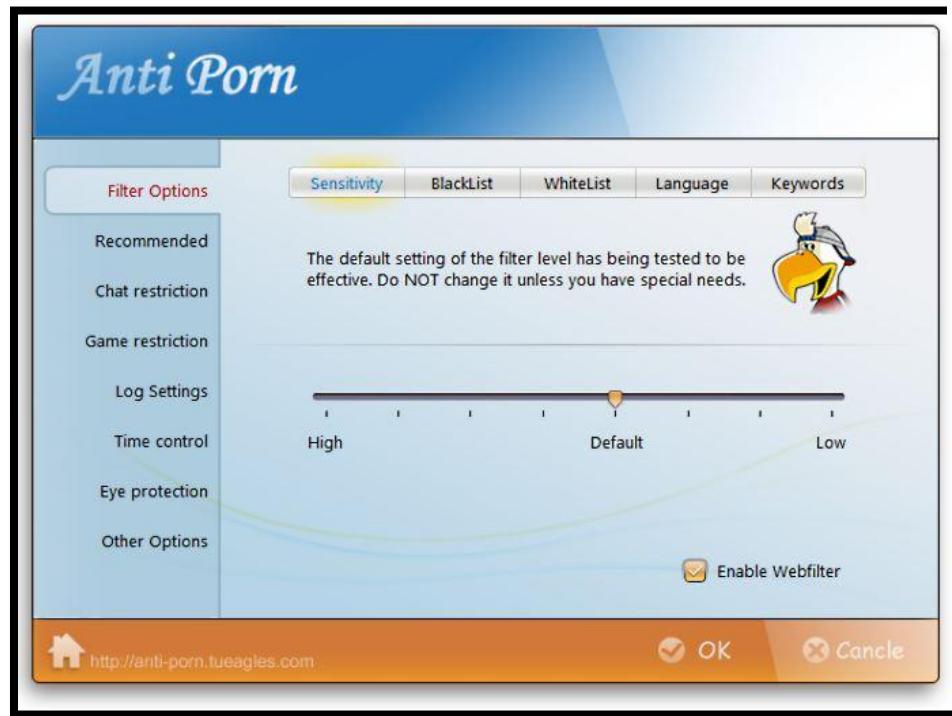


الشكل 13 صورة توضح واجهة We-Blocker

- يتم تسجيل نشاط المستخدم على الإنترنت (الموقع التي تم دخلوها)؛

- يتم حصر دخول المستخدم إلى المواقع على الإنترنط ما عدا تلك المخزنة بقائمة المواقع المسموح بزيارتها والمسماه بالقائمة البيضاء؛
- يتم تصنيف المواقع على حسب محتوياتها وبالتالي عند وضع علامة صح أمام تصنيف منها فيتم حجب الدخول إلى المواقع التي تدرج تحت هذا التصنيف؛
- يتم حجب المواقع التي تحتوي على كلمات سببية غير مرغوب فيها؛
- الإقلال من وضع حدود على استخدام الإنترنط لأن ذلك سيقلل من سرعة تحميل الصفحات بالإضافة إلى صعوبة الدخول إلى المواقع في بعض الأحيان.

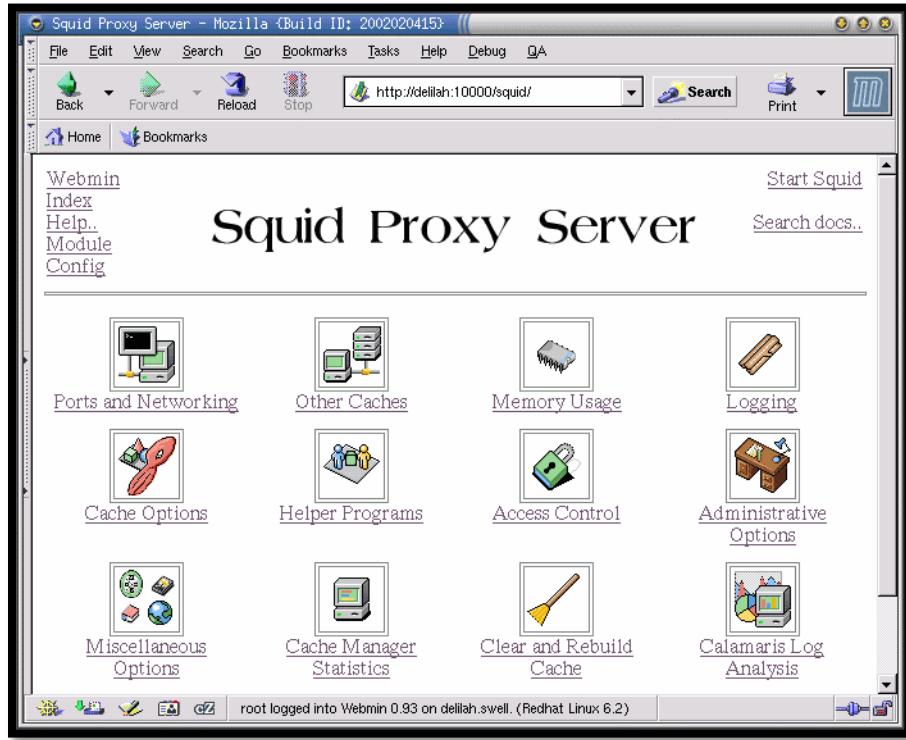
### (Windows 2000/2003/XP/Vista/7) Anti-Porn 1.5



الشكل 14 صورة توضح واجهة anti-porn

- حجب المواقع الغير مرغوب فيها وبرامج الشات؛
- الإغلاق قبل العرض، يقوم Anti-Porn بالإطلاع السريع على كل محتويات الصفحة قبل فتحها ويغلق الصفحات المنوعة؛
- إمكانية إضافة أي موقع إلى القائمة السوداء ليتم حجبها؛
- إمكانية استثناء أي موقع قام البرنامج بحجبه من خلال إضافة رابط الموقع إلى القائمة البيضاء؛
- إمكانية تحديد اللغات التي سيتعامل معها البرنامج للتحكم في فلترتها؛
- إمكانية إضافة أي كلمة إلى قائمة الكلمات المفتاحية ليقوم البرنامج بحجب المواقع المتضمنة لهذه الكلمة تلقائياً؛
- تحديد الدردشة على الإنترنت، يمكن لـ Anti-Porn تحديد استخدام برامج شائعة للدردشة كما يمكنه تحديد الدخول الى غرف دردشة أساسية؛
- تحديد أوقات الدخول إلى الإنترنت باليوم والساعة؛
- سجل التاريخ، يحتفظ Anti-Porn بسجل كامل لكل المواقع التي تم زيارتها سواء تم تنقيتها أم لا؛
- ينصح بالموقع الموثوق فيها، Anti-Porn يقوم بترشيح عدد من المواقع المناسبة مع إمكانية إضافة المزيد من المواقع التي تريدها؛
- مجاني ومفتوح المصدر.

## Squid 1.6



الشكل 15 صورة توضح واجهة squid

- خادم بروكسي يعمل مع أنظمة تشغيل Unix و Linux و حالياً قامت شركة بإعداده للعمل ضمن الويندوز؛
- مستخدم من قبل مئات من مزودي خدمة الإنترنت لتقديم مستخدميه بأفضل الطرق للوصول إلى الإنترت؛
- يقوم بمراقبة المطبيات المنقولة بين المخدم والزيون والعمل على تحسين الأداء وتخزين الصفحات التي تم زيارتها مؤخراً
- والأكثر زيارة مع تحديد المستخدم والوقت والتاريخ بدقة؛
- مسرع إنترنت باستخدام خاصية التخزين المؤقت caching وفقاً للبروتوكولات HTTP, HTTPS, FTP،
- Gopher، يقوم بتوزيع الإنترت على الأجهزة التابعة للشبكة حسب الإعدادات الخاصة فيه؛

- حجب المواقع التي لا تريد أن تتصفحها أجهزة الشبكة؛

- فلترة المواقع والكلمات وتحديد سماحيات الوصول؛

- مجاني ومفتوح المصدر.<sup>[4]</sup>

## SquidGuard 1.7

- هو إضافة squid على add-on،

- يقوم بعمل حجب للمواقع التي لا تريد أن يتم تصفحها عبر أجهزة الشبكة؛

- يحتوي على قاعدة بيانات تحتوي على المواقع "السوداء"؛

- إمكانية عمل صلاحيات للمستخدمين بطريقة أفضل وأسهل من squid؛

- الحد من الوصول إلى شبكة الإنترنت البعض المستخدمين إلى قائمة خوادم الويب المعروفة والمقبولة و / أو عناوين URL

فقط؛

- منع الوصول إلى بعض خوادم الويب المدرجة في القائمة السوداء و / أو عناوين URL لبعض المستخدمين؛

- منع الوصول إلى عناوين URL تحوي كلمات غير مرغوب بها لبعض المستخدمين؛

- إمكانية استخدام التعبيرات النظمية لحجب أجزاء محددة من صفحة الويب قبل عرضها؛

- فرض استخدام domain names / حظر استخدام عنوان IP في عناوين URL

- إعادة توجيه عناوين URL المحظورة إلى صفحة أخرى؛

- إعادة توجيه مستخدم غير مسجل لاستماراة التسجيل؛

- إعادة توجيه الشرائط الإعلانية إلى صورة فارغة؛

- إمكانية توليد ملفات log files؛

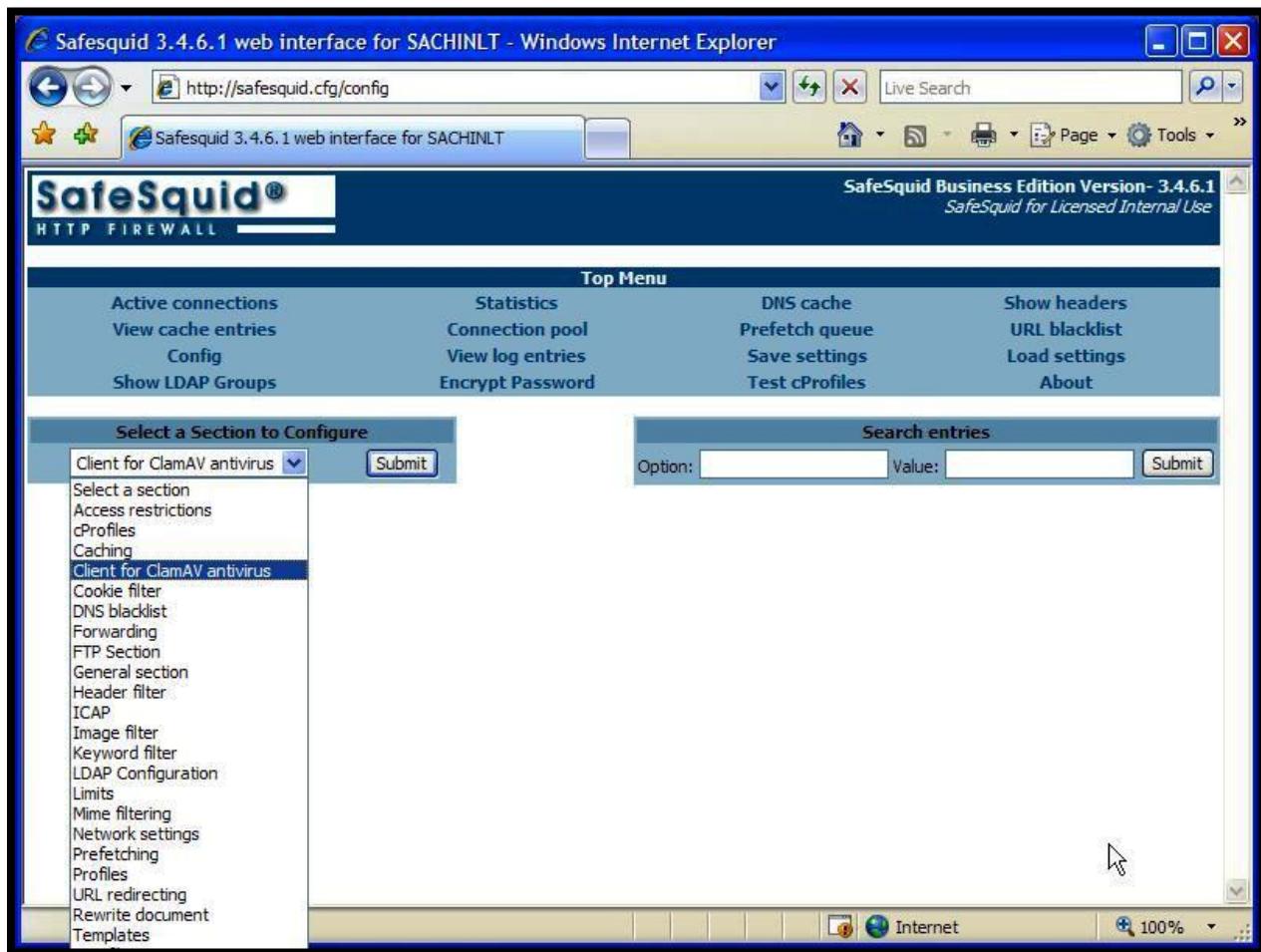
- لديها قواعد وصول اعتماداً على الوقت من اليوم، يوم من الأسبوع، والتاريخ ...؛

- لديها قواعد مختلفة ل مختلف مجموعات المستخدمين (Profiles/Grouping)؛

- مجاني، سريع ، سهل التنصيب ، سهل الاستخدام و مفتوح المصدر؛

- من أنظمة التشغيل التي تعمل معها:

- **AIX: 4.1.3, 4.3.2.0/egcs-2.91.66**
- **Dec-Unix: OSF1-4.0/gcc-2.7.2.3, 3.2C/gcc-2.7.2.3**
- **FreeBSD 4.x-STABLE gcc 2.95.3**
- **Linux: RedHat-5.2/gcc-2.8.1 RedHat-7.x/gcc-2.8.1, Gentoo 1.12.6/gcc-3.3.6**
- **Solaris: 2.6/gcc-2.7.2.3 2.6/gcc-2.95.3, 2.8/gcc-2.95.3**
- **CentOS: 4.4**



الشكل 16 صورة توضح واجهة safe-squid

- هو إضافة squid على add-on ،

- يقوم بعمل حجب للموقع التي لا تريد أن يتم تصفحها عبر أجهزة الشبكة ،

- يحتوي على قاعدة بيانات تحتوي على الموقع "السوداء" ،

- توفر خدمة ال caching ،

- توفر خدمة مراقبة المستخدمين؛
- ضبط عرض الحزمة bandwidth؛
- مستخدم من قبل المدارس والشركات ومزودي خدمة الإنترنت؛
- منتج برمجي تجاري، مجاني فقط في حال لدينا ثلاثة مستخدمين أو أقل ومفتوح المصدر؛
- يؤمن واجهة تخطيطية لتسهيل عملية الاستخدام من قبل المدير ليقوم بوضع إعداداته الخاصة؛
- إمكانية حجب موقع كاملة أو حجب الموقع بسبب احتواء ال URL الخاص فيه على كلمات محددة مسبقاً؛
- إمكانية حجب الموقع اعتماداً على كلمات محددة؛
- إمكانية حجب أنواع محددة من الملفات مثل ملفات الصوت والصورة؛
- إمكانية استخدام التعابير النظمية لحجب أجزاء محددة من صفحة الويب قبل عرضها؛
- إمكانية إخفاء الشرائط والصفحات الإعلانية لتخفيف عرض الحزمة وبالتالي تحسين الأداء؛
- إمكانية حجب الصور السيئة الغير المرغوب فيها من خلال تحليل الصور والتعرف على محتواها؛
- يولد ملفات logs لمراقبة المستخدمين ومن ثم ترجمة هذه الملفات وتوليد تقارير تحليلية؛
- يملك خاصية الملف الشخصي والذي يتتيح للمدير أن يصنف المستخدمين ضمن مجموعات ووضع إعدادات الفلترة على مستوى المجموعات؛
- دقتها تراوح بين 85%-90% وتعتبر رادع جيد؛
- فضلاً عن أنها تعتبر مضاد للفيروسات والبرمجيات الخبيثة؛
- أنظمة التشغيل التي تعمل معها [7].Microsoft Windows, Linux -

- مجاني ومفتوح المصدر؛

- يدعم أنظمة التشغيل التالية:

Linux distributions (RedHat, CentOS, Fedora, Debian, Ubuntu),

Windows

- يعمل كإضافة على مخدم البروكسي Squid [5]؛

- يقوم QuintoLabs بالفلترة بواسطة عدة تقنيات هي:

1- يقوم بضمان فلترة الموقع المُزارة (من موقع البالغين) من خلال تقنية RTL Labeling وهي عملية وضع

علامات خاصة مثل HTTP headers أو meta بالصفحات والصور التابعة لموقع البالغين، تتم هذه

العملية من قبل الموقع طوعاً، فعندما يقوم QuintoLabs بفحص الصفحات يبحث عن هذه العلامات فإذا

ووجدها يقوم بحجب الصفحة، تقنية RTL ليس لها أي إعدادات خاصة وإنما فقط يمكن تفعيلها أو إلغاء

تفعييلها on/off؛

2- URL-Heuristics يتم استخدام هذا module لحجب URLs معينة التي غالباً ما تكون مرتبطة

بموقع البالغين، وهنا يوجد عدة مستويات من الصرامة في الاستدلال تتبعها الخوارزمية مثلاً

heuristics\_level = high ، إذا كان مستوى الصرامة عالي جداً فإننا قد نحصل على بعض الإجابات

الخاطئة بمعنى أنه قد نحصل على نتيجة بأن URL ما هو تابع لموقع له علاقة بالبالغين رغم أنه لا يكون

فعلاً كذلك !؛

3- Full Text Content Inspection وهي عملية الفحص الكامل للنص الموجود في الصفحات

وأغراض HTML, Json، حيث يتم تحديد الكلمات والعبارات المراد حجبها ضمن ملفات محددة، وعندما

يقوم التطبيق بفحص صفحة ما يقوم بالبحث عن هذه الكلمات وحساب وزن الصفحة اعتماداً عليها فإذا تجاوز

هذا الوزن حداً معيناً يتم حجب الصفحة، تعد هذه العملية مكلفة، لذلك تم توفير إمكانية تحديد الحجم الأقصى

لملف الذي سيطبق عليه عملية الفحص الكامل لتجنب فحص الملفات الكبيرة التي سيأخذ فحصها وقتاً كبيراً،

الحجم الافتراضي 300k

4- إمكانية حجب الموقع فوراً من خلال اسمه أو عنوانه URL وذلك من خلال كتابتها في ملفات خاصة:

مثلاً حجب موقع حسب العنوان:

url = http://.\*sex\com.\*

حجب موقع حسب الاسم:

domain = www.porno.com

حجب جميع المسارات الجزئية subdomains في الموقع السابق:

domain = .porno.com

كما يمكن حجب الموقع من خلال وضعها ضمن تصنيفات معينة ومن ثم حجب التصنيف، ويمكن إضافة التصنيف بالشكل التالي:

[6].block\_category = sexuality

### Mثال عن آلية عمل أحد الفلاتر وهو DansGuardian

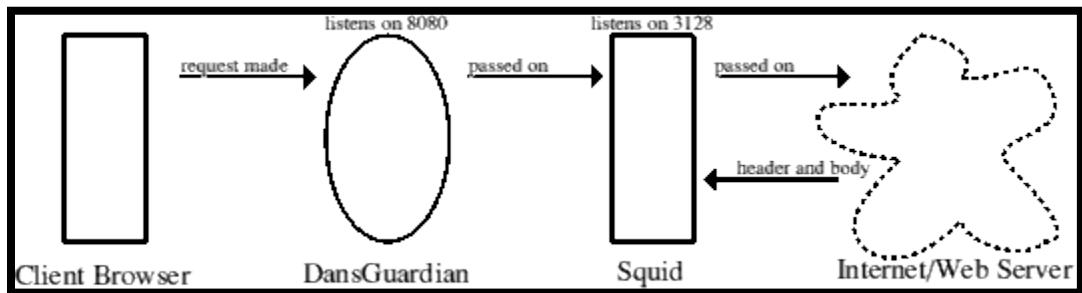
- يرسل المستخدم الطلب إلى DansGuardian، يقوم DansGuardian بفحص ترويسة الطلب القادم من

المستخدم آخذاً بعين الاعتبار اسم المستخدم، IP، URL، المرسل،

- يتم تطبيق الفلاتر المناسبة، banned user, exception user, banned URL, exception URL

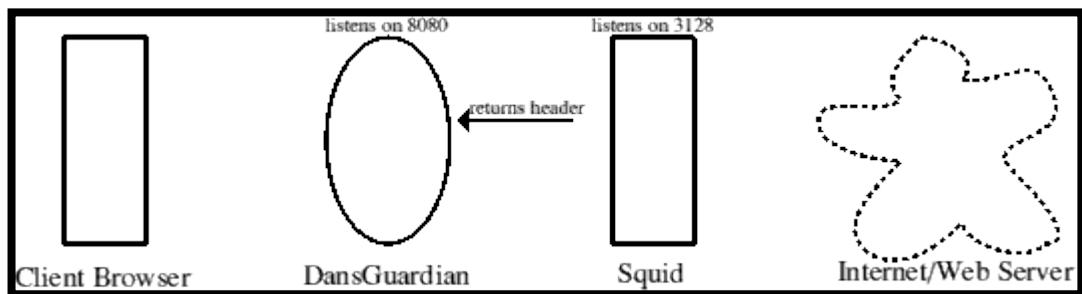
banned IP, exception IP

- في حال كانت النتيجة "قبول" يتم تمرير الطلب إلى البروكسي squid والذي يقوم بإحضار الملف المطلوب من الإنترنت،



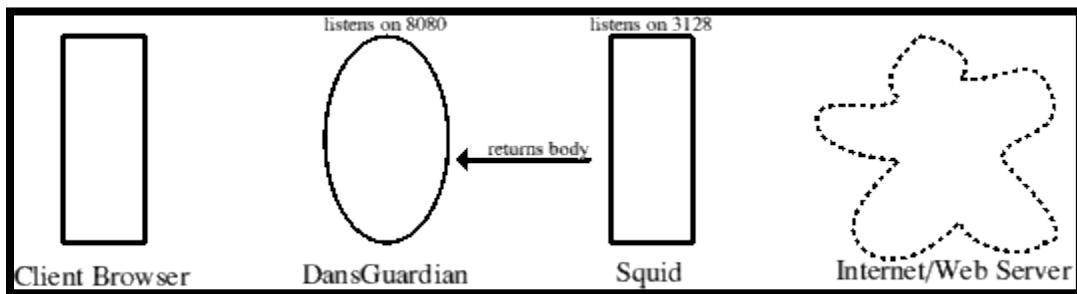
**الشكل 17 آلية العمل مرحلة 1 DansGuardian + squid**

- يقوم squid بتمرير ترويسة الملف فقط إلى DansGuardian، والذي بدوره يقوم بفحص الترويسة بتطبيق مجموعة من الفلاتر المناسبة banned MIME-types وامتداد الملف؛
- في حال كانت النتيجة “قبول” يتم الانتقال إلى الخطوة التالية؛



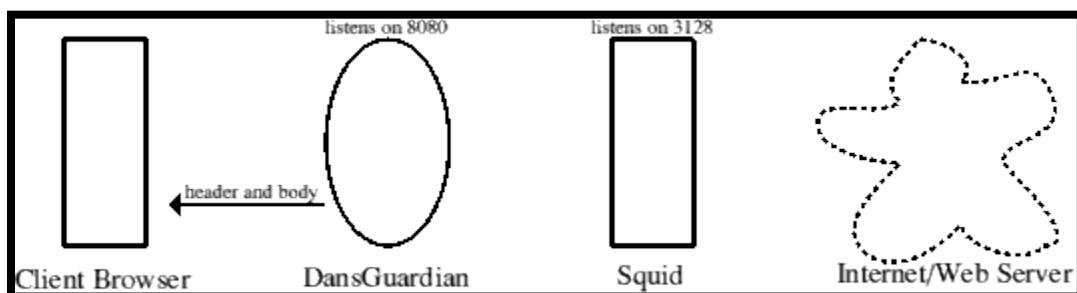
**الشكل 18 آلية العمل مرحلة 2 DansGuardian + squid**

- يقوم squid بتمرير محتوى الملف أو الوثيقة إلى DansGuardian، والذي بدوره يقوم بتقسيمه إلى مجموعة عبارات ومن ثم يقوم بتطبيق الفلاتر المناسبة banned phrase, exception phrase؛



الشكل 19 آلية العمل DansGuardian + squid مرحلة 3

- في حال كانت النتيجة "قبول" يقوم DansGuardian بتمرير ترويسة ومحظى الملف إلى متصفح المستخدم الطالب.



الشكل 20 آلية العمل DansGuardian + squid مرحلة 4

<http://lifehacker.com> تصویت من قبل موقع

#### Which Content Filtering Tool is Best?

- **OpenDNS 57%**
- 
- **K9 12%**
- **Hosts File 10%**
- **DansGuardian 8%**
- **Other 8%**
- **SquidGuard/Squid 5%**

#### Hosts File 1.1.1

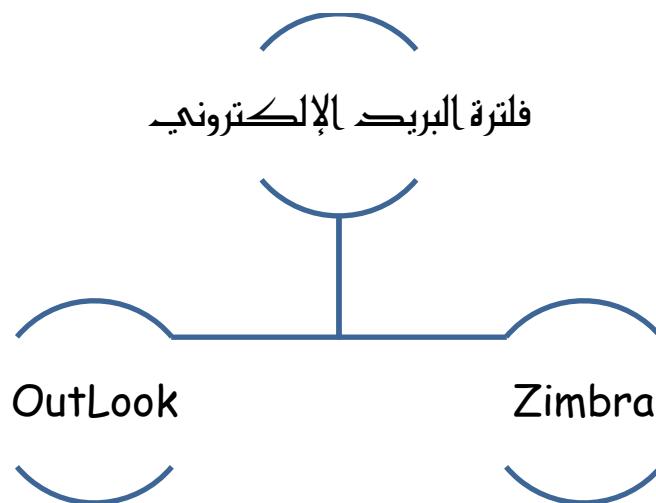
- ملف شبيه بدليل الهاتف يحوي المواقع المطلوب حجبها مع ال **ips** المقابلة لها؛
- عند محاولة الدخول إلى أحدى هذه المواقع عن طريق كتابة ال **URL** أو ال **IP** المقابل وكان ذلك العنوان موجود ضمن الملف على أنه موقع محظى سيتم إظهار رسالة خطأ وإنلا سيتم تحميل الصفحة الموافقة للعنوان المطلوب.

## الفصل الثالث

### دراسة مرجعية عن فلترة وتصنيف البريد

#### الإلكتروني

نعرض في هذا الفصل دراسة مرجعية عن كيفية إدارة outlook و برنامج zimbra للبريد الإلكتروني، وآلية إضافة القواعد المطبقة على البريد الإلكتروني عند وروده بغية حجبه أو وضعه في مجلد خاص وغير ذلك من العمليات التي من الممكن تطبيقها على البريد الإلكتروني الوارد أو المرسل.



الشكل 21 النقاط الأساسية في فلترة البريد الإلكتروني

## ١ إدارة البريد الإلكتروني من خلال القواعد Outlook

- القاعدة هي حدث يتم تطبيقه على البريد الإلكتروني الوارد أو المرسل اعتماداً على شروط محددة تم تعريفها مسبقاً ضمن القاعدة.

- يمكن اختيار أكثر من شرط ضمن القاعدة وهذا ما يسمى .Meta Rules

- تتوزع القواعد ضمن مجموعتين مجموعه تدعى Organization و مجموعة تدعى .Notification

- هناك أنواع للقواعد:

-١ Stay Organized فمثلاً يمكن وضع قاعدة للرسائل الواردة من مرسل معين مثل Bobby Moore والتي

. تحوي الكلمة sales ضمن العنوان وعند ورود رسالة تتحقق هذه القاعدة يتم وضعها في مجلد Bobby's Sales

-٢ Stay Up to Date هذه القواعد تنبئ المستقبل عند وصول رسالة خاصة فمثلاً يمكن وضع قاعدة لإرسال تنبيه

على الموبايل عند ورود بريد إلكتروني من أحد أفراد العائلة.

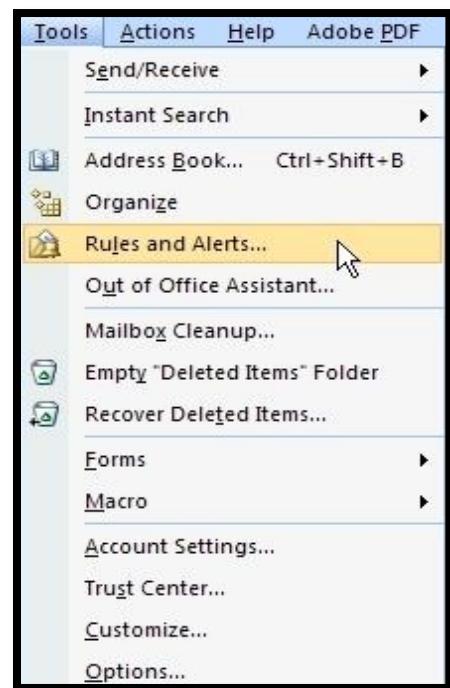
-٣ Start from a blank rule يتم بناء القواعد هنا من دون الاستعانة ب .rule template

-٤ يتضمن templates لقواعد Outlook يمكن استخدام هذه النماذج لبناء القواعد أو أن يتم تصميم القواعد بطريقة OutLook واضع القواعد.

## 1.10 استخدام Outlook rule

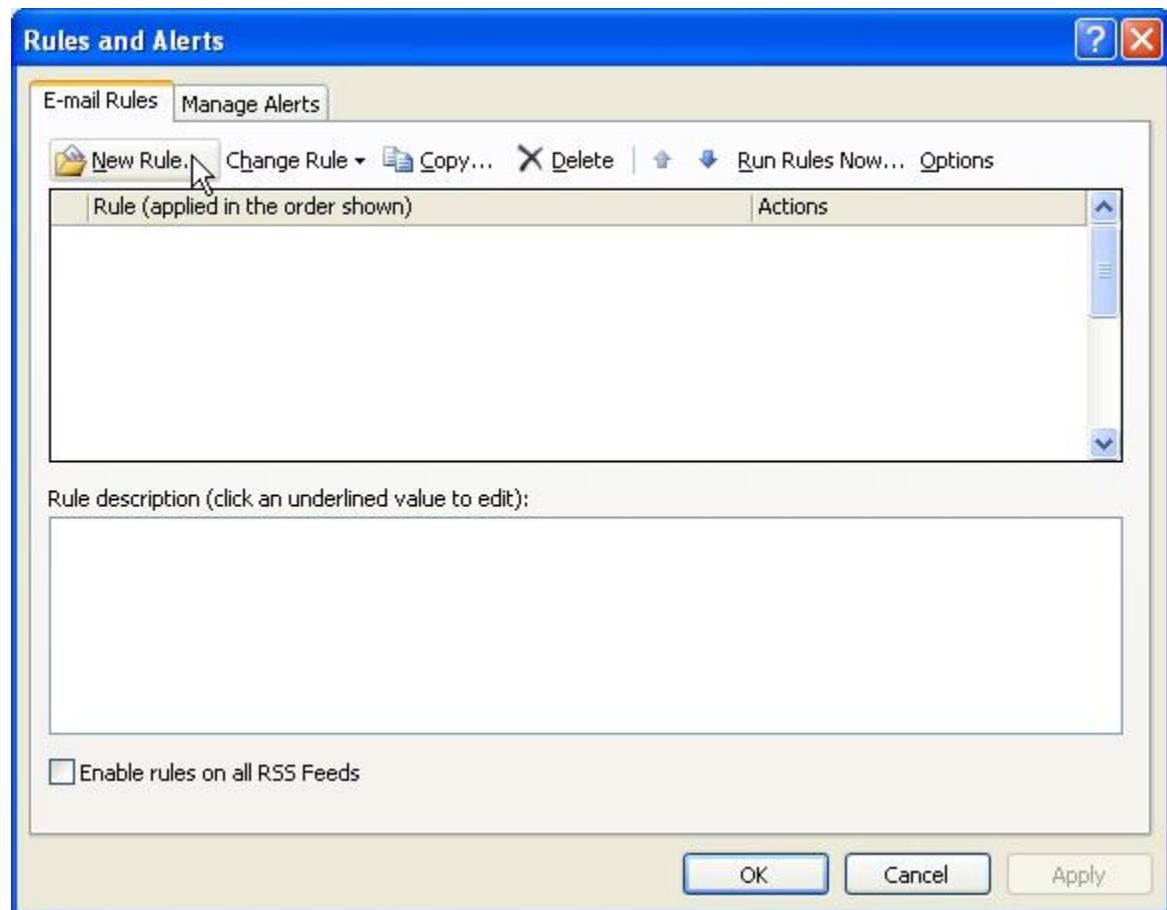
1- يتم اختيار القائمة file

2- يتم اختيار القائمة Manage Rules & Alerts



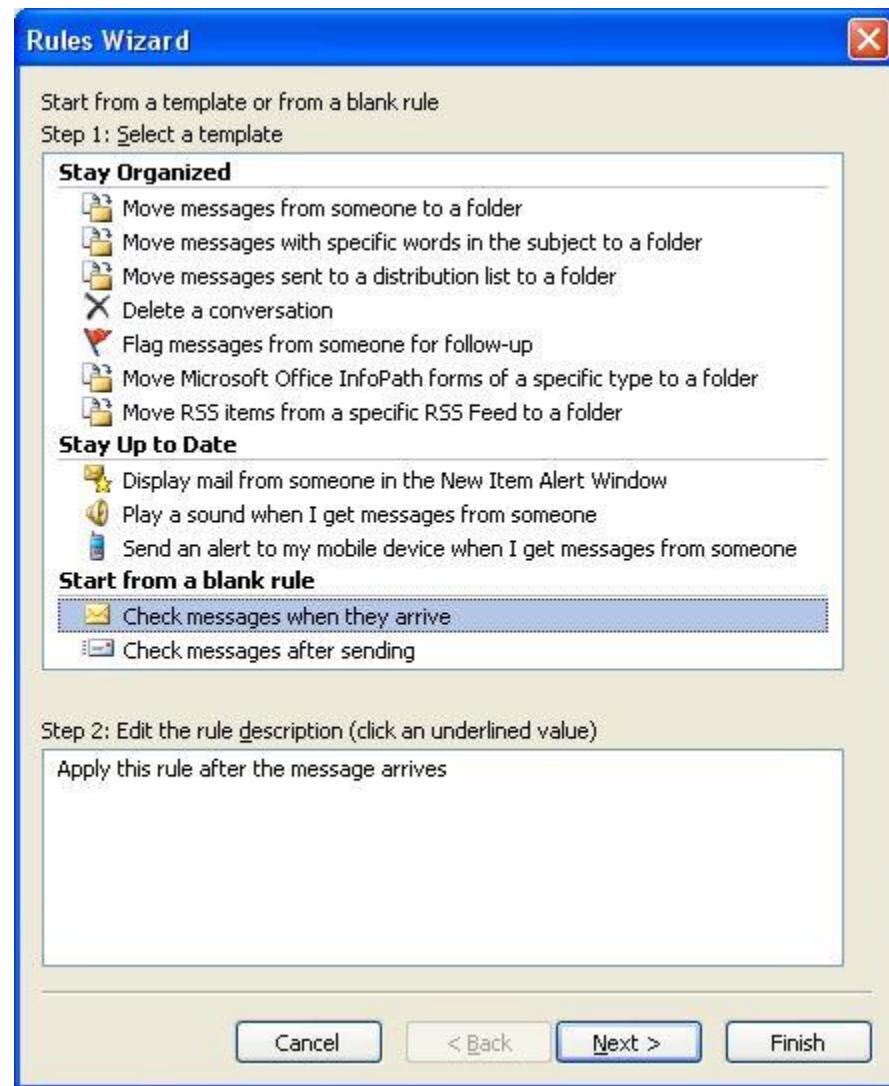
الشكل 22 واجهة outlook مرحلة

-3 من New Rule يتم اختيار القائمة E-mail Rules ومن ثم اختيار Rules and Alerts



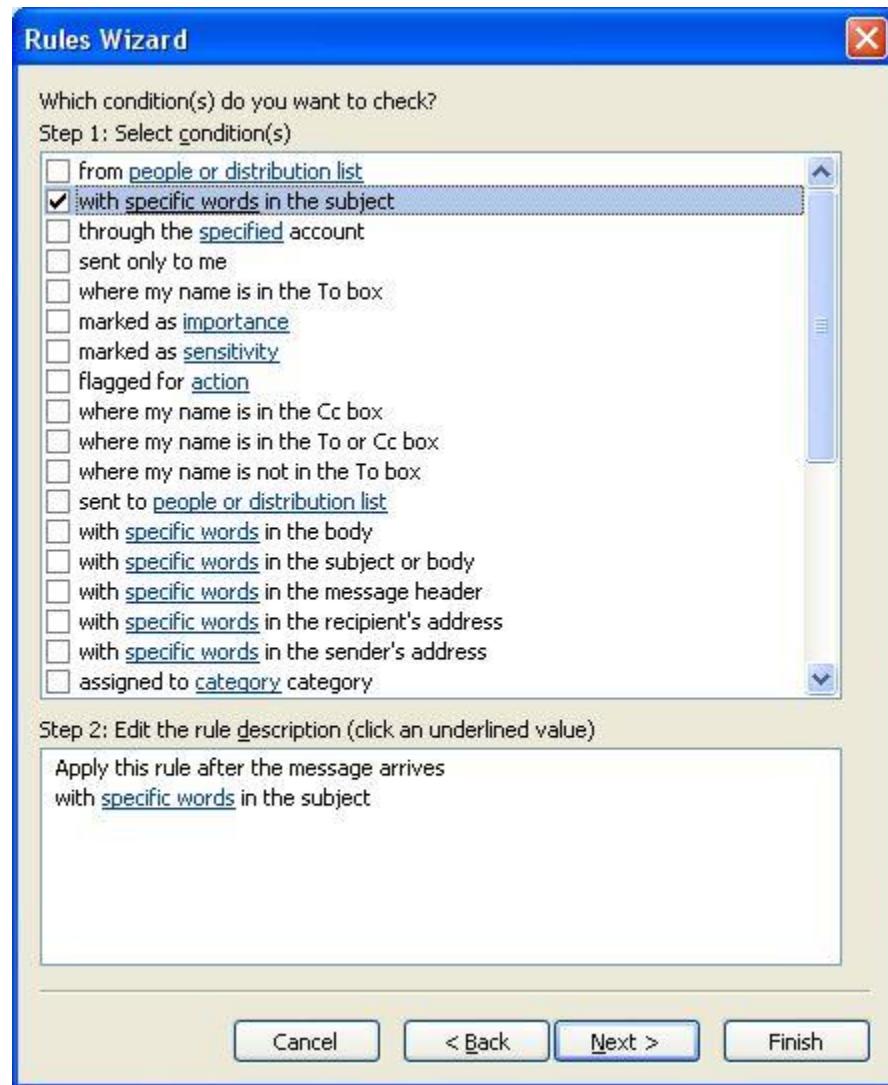
الشكل 23 واجهة outlook مرحلة 2

4- بعد ذلك تظهر نافذة Rules Wizard وتحتار قالب محدد أو بناء القاعدة من دون الاعتماد على القوالب الموجودة مسبقاً، وبفرض أنه تم اختيار قالب Start from a blank rule وتحديداً فحص البريد الإلكتروني عند وروده، هذا ويمكن تعديل توصيف القاعدة.



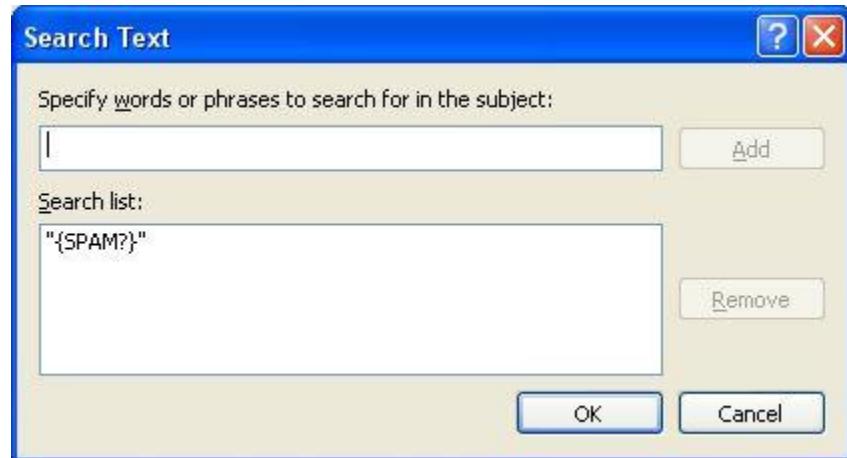
الشكل 24 واجهة outlook مرحلة 3

5- يتم اختيار مجموعة الشروط التي يتم تطبيقها على الرسالة على سبيل المثال فلترة البريد الإلكتروني بناء على كلمات محددة في موضوع الرسائلن كما يمكن تعديل التوصيف كما هو مبين في الشكل.

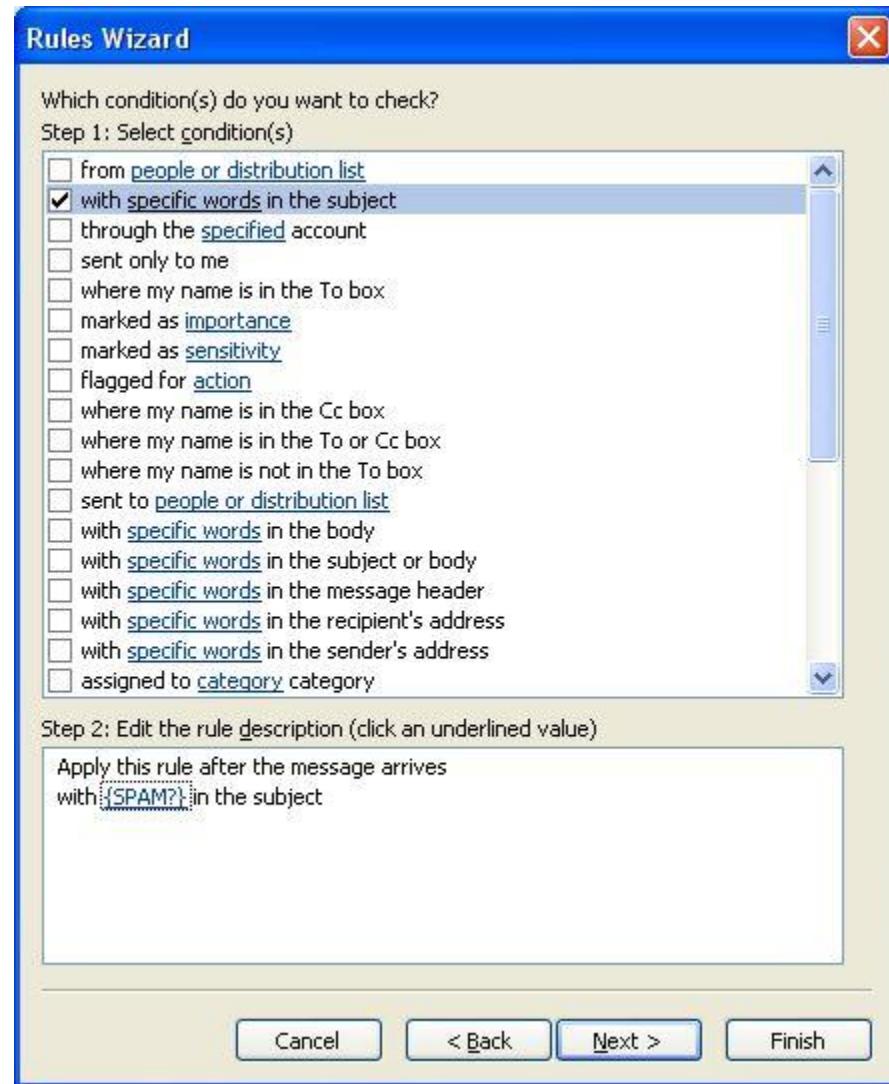


الشكل 25 واجهة outlook مرحلة 4

6- تحديد الكلمات الغير مرغوب بها والتي ستتم الفاترة على أساسها



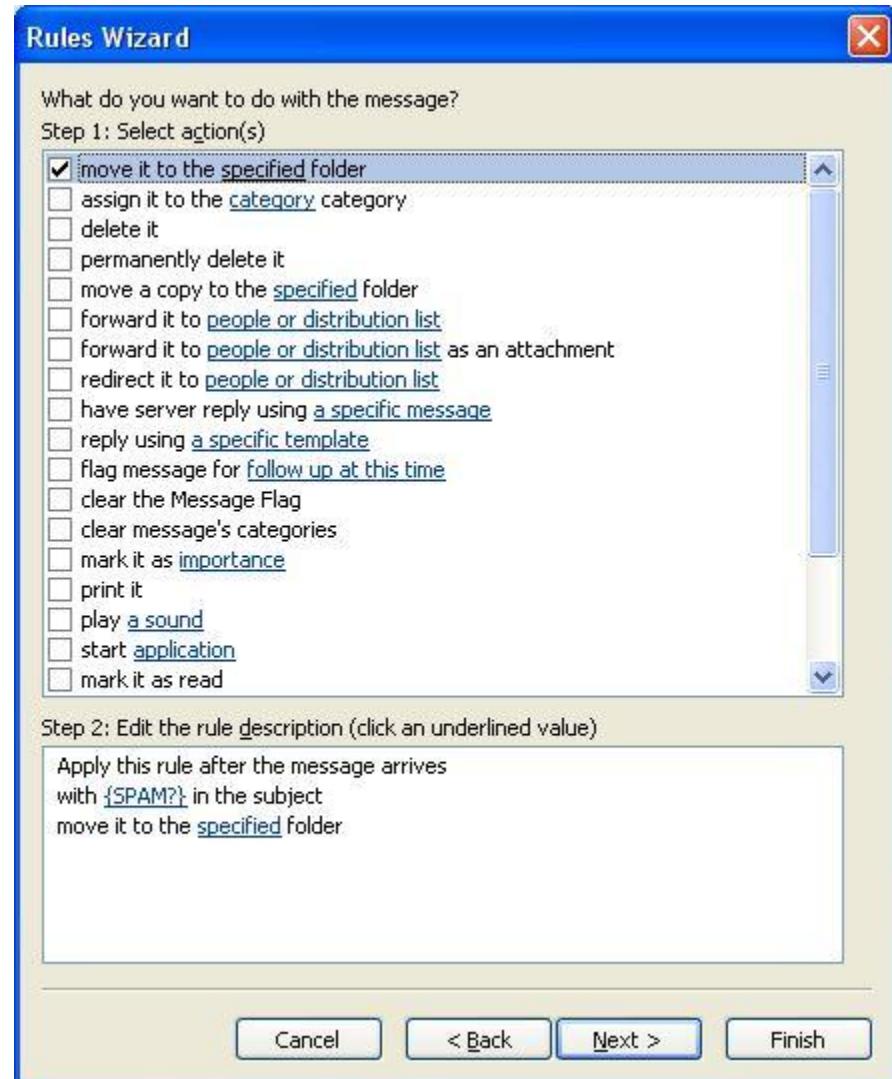
الشكل 26 واجهة outlook مرحلة 5



الشكل 27 واجهة outlook مرحلة 6

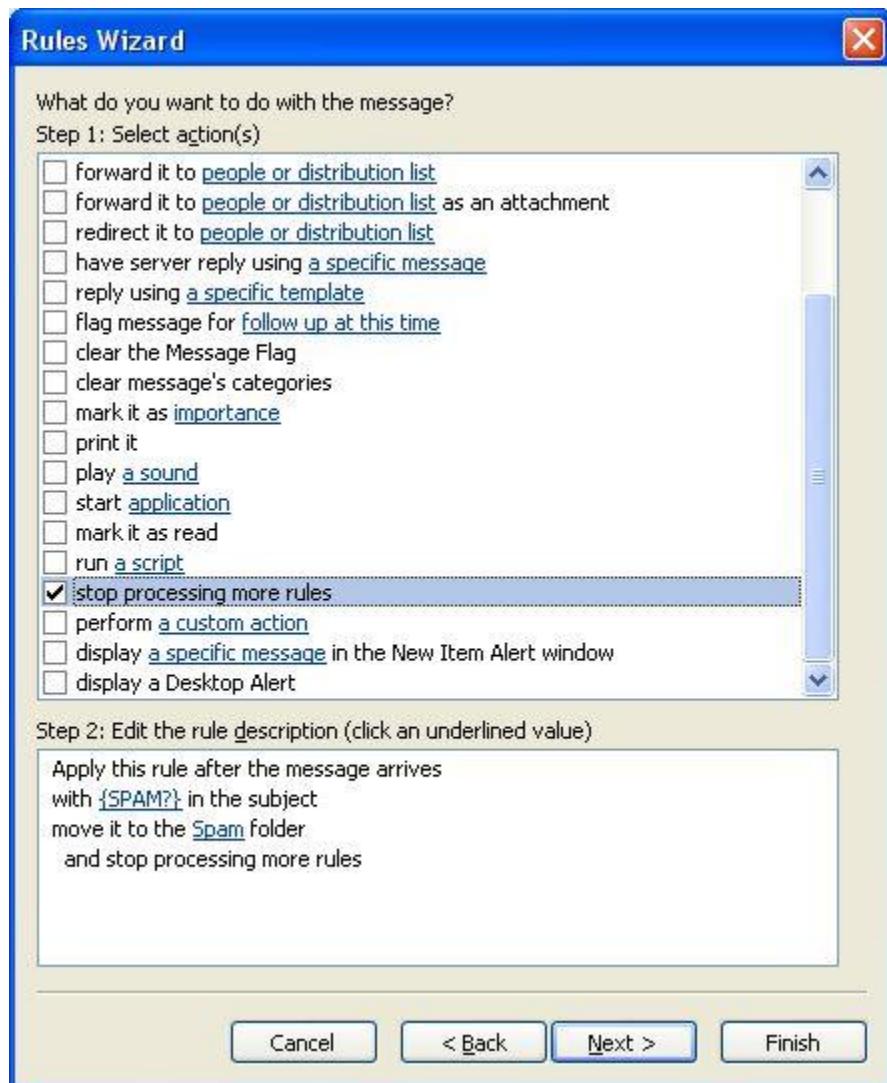
7- نقل البريد الإلكتروني الذي يحقق هذه القاعدة إلى مجلد خاص وليكن spam أو تفعيل خيار حذف هذه الرسالة في حال

ثبت أنها spam



الشكل 28 واجهة outlook مرحلة 7

– تفعيل خيار إيقاف فحص بقية القواعد عند تحقق هذه القاعدة



الشكل 29 واجهة outlook مرحلة 8

- Zimbra هو نظام Anti-spam يستخدم لفلترة البريد الإلكتروني؛
- الطريقة الأسهل هو التعديل على ملف /opt/zimbra/conf/salocal.cf.in؛
- الطريقة الأبسط المتبعة في الفلترة هي القوائم البيضاء والقوائم السوداء؛
- القوائم السوداء تحجب البريد الإلكتروني الوارد من عنوان أو domain محدد؛
- القوائم البيضاء تسمح بعرض البريد الإلكتروني الوارد من عنوان أو domain محدد؛
- لإضافة القوائم السوداء أو البيضاء يتم الدخول إلى الملف salocal.cf.in والكتابة بالصيغة التالية على سبيل المثال:

```
blacklist_from sales@traveloforrange.com
```

```
whitelist_from bill@yahoo.net
```

```
blacklist_from *@emn-mysavingsnow.net
```

- حيث يتم حجب كل المستخدمين التابعين لل domain التالي emn-mysavingsnow.net؛
- عند الانتهاء من التعديل على الملف يتم إعادة تشغيل zimbra من خلال التعليمية التالية:

**zmmtactl restart && zmamavisdctl restart**

- يتم قراءة عنوان ومحظى البريد الإلكتروني وتطبيق القواعد عليهما؛
- القواعد هنا قد تحوي كلمات بعینها أو تعابير نظامية؛

- عندما يتحقق البريد الإلكتروني القاعدة الموضعة يتم زيادة الرصيد الكلي للبريد الإلكتروني بمقدار رصيد هذه القاعدة؛
- عندما يتجاوز الرصيد الكلي للبريد الإلكتروني عتبة محددة يتم اعتباره spam ، وإذا كان أكبر بمقدار معين يتم حذفه مباشرةً؛
- يتم إضافة هذه القواعد الحاوية على التعابير النظمية إلى الملف alocal.cf.in بالشكل التالي :

```
body LOCAL_RULE /sale/
score LOCAL_RULE 0.5
```

- القاعدة السابقة اسمها LOCAL\_RULE تعني إذا احتوى البريد الإلكتروني على كلمة sale سيتم زيادة الرصيد الكلي للبريد الإلكتروني بمقدار 0.5 ؛
- Meta rule هي قاعدة تحوي على مجموعة قواعد يجب أن تتحقق معاً حتى يتم زيادة الرصيد الكلي للبريد الإلكتروني بمقدار معين مثال :

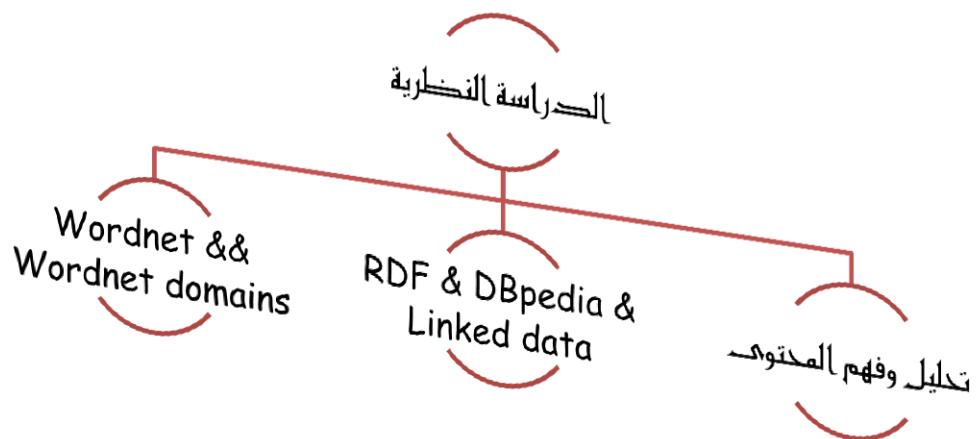
```
body LOCAL_FOUR_CAPS / [A-Z] [A-Z] [A-Z] [A-Z] /
body LOCAL_MONEY /\d?\d?\d?.\d\d\b/
meta LOCAL_STOCK (LOCAL_MONEY && LOCAL_FOUR_CAPS)
score LOCAL_STOCK 1
```

- بمعنى إذا حق البريد الإلكتروني القاعدة LOCAL\_FOUR\_CAPS والقاعدة LOCAL\_MONEY . سيتم زيادة الرصيد الكلي للبريد بمقدار 1.



## الباب الثاني

### الدراسة النظرية



الشكل 30 النقاط الأساسية في مرحلة الدراسة النظرية



# الفصل الرابع

## Wordnet && Wordnet Domains

---

نعرض في هذا الفصل لمحة عن معجم wordnet حيث تم الاعتماد عليه في فهم المحتوى، كما نعرض مجالات هذا المعجم والتي يتم الاعتماد عليها في تحديد أهم المجالات التي تتحدث عنها الصفحة ونسبة انتقاء كل مجال إلى الصفحة.

### 1 معجم wordnet3

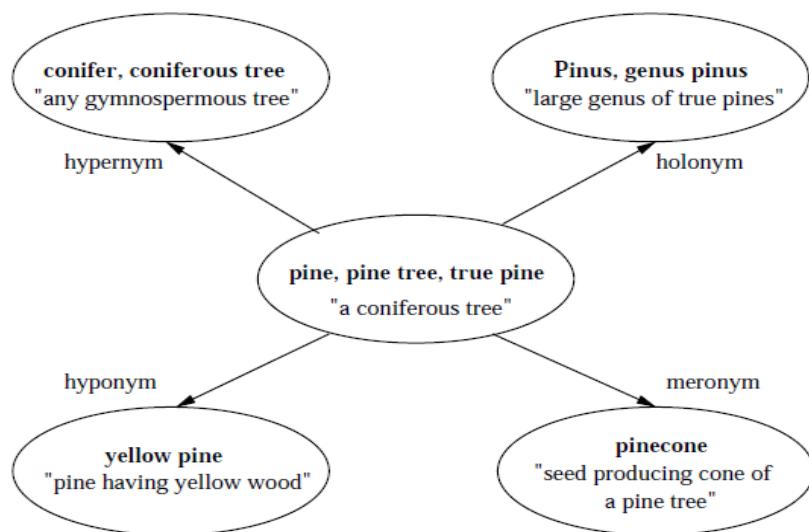
معجم الـ WordNet عبارة عن معجم يضم كلمات اللغة الانكليزية، بالإضافة إلى شرح هذه الكلمات المجمعة في مجموعات (Synsets)، بحيث أن كلمات كل مجموعة لها نفس المعنى (synonyms)، وهناك مجموعة من العلاقات الدلالية (semantic) التي تربط هذه المجموعات.

من المعلوم أن بعض الكلمات لها معنى وحيد (monosemous) و الكثير من الكلمات تحمل أكثر من معنى واحد (polysemous) وذلك حسب سياق الحديث، فعلى سبيل المثال كلمة "wristwatch" ليس لها إلا معنى وحيد، وبالتالي فهي تنتمي إلى مجموعة واحدة، في حين كلمة "bark" تحمل أكثر من معنى، فمن الممكن أن تأتي بمعنى لحاء الشجر، أو بمعنى صوت الكلب، وبالتالي فهي تنتمي إلى أكثر من مجموعة واحدة . [16]

في WordNet يتم وضع الكلمات المترادفة في مجموعات تدعى (synsets)، وكل مجموعة تفسير (Gloss) وهو عبارة عن جملة تشرح المفهوم الذي تمثله هذه المجموعة، فعلى سبيل المثال المجموعة التالية : { *foundation* }, *base,basis* ، لها التفسير: “lowest support of a structure” ، *cornerstone, groundwork,fundament*

وهذه المجموعات ترتبط مع بعضها البعض عن طريق علاقات دلالية وهي: meronym ، hypernym ، hyponym . holonym ،

مثال على العلاقات المختلفة التي تربط بين مجموعات الكلمات في الـ WordNet :



الشكل ٣١ مثال توضيحي عن العلاقات الدلالية في wordnet

أمثلة توضيحية عن العلاقات الدلالية المختلفة :

**Hypernyms:** Y is a hypernym of X if every X is a (kind of) Y (*canine is a hypernym of dog*, because every dog is a member of the larger category of canines).

**Hyponyms:** Y is a hyponym of X if every Y is a (kind of) X (*dog is a hyponym of canine*). --

**Holonym:**  $\gamma$  is a holonym of  $X$  if  $X$  is a part of  $\gamma$  (*building is a holonym of window*).

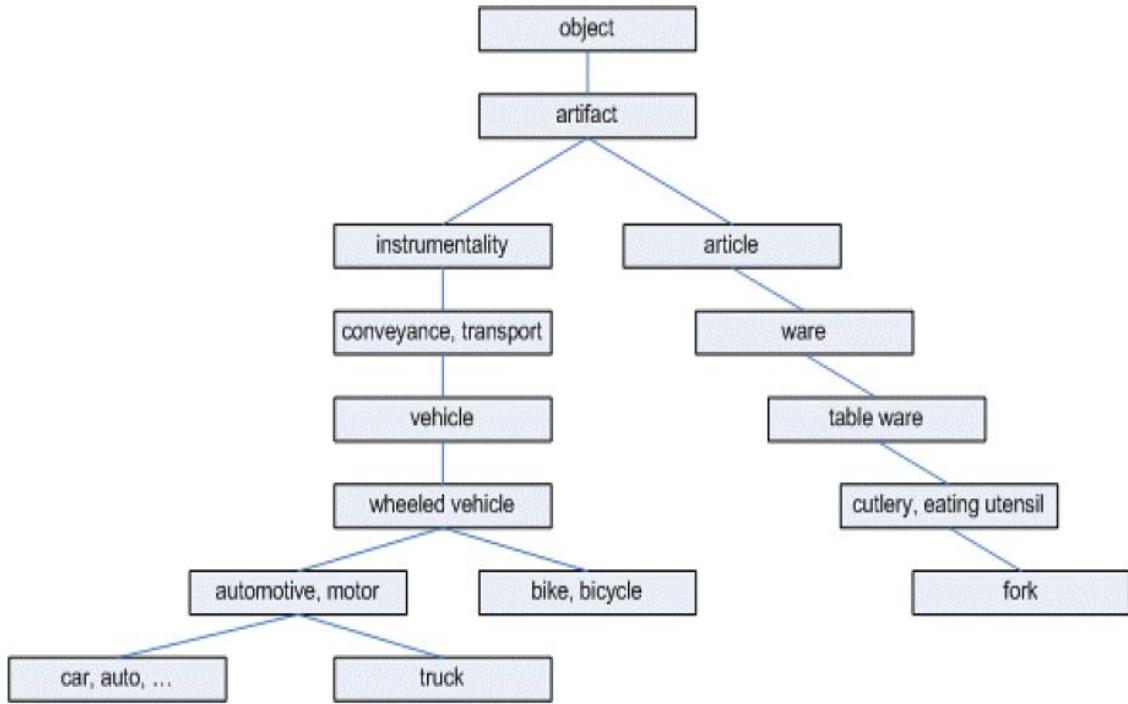
**Meronym:**  $\gamma$  is a meronym of  $X$  if  $\gamma$  is a part of  $X$  (*window is a meronym of building*).

**Hypernym:** the verb  $\gamma$  is a hypernym of the verb  $X$  if the activity  $X$  is a (kind of)  $\gamma$  (*to perceive is an hypernym of to listen*).

**Troponym:** the verb  $\gamma$  is a troponym of the verb  $X$  if the activity  $\gamma$  is doing  $X$  in some manner (*to lisp is a troponym of to talk*).

ولكننا في دراستنا هذه سوف نهتم بالعلاقة (is a kind of) hypernymy و العلاقة المعاكسة (is kind of) hyponymy . كما أن الكلمات بنية هرمية باستخدام علاقتي (hyponym و hypernym ) والتي نستطيع من خلالها إيجاد المفهوم (concept) لكلمة ما .

مثال: نلاحظ من الشكل التالي أن الـ "car" هي نوع من الـ "vehicle" ، كما وتنبيح هذه البنية معرفة متزادات الكلمة car motorcar، machine، automobile، . auto وهي



**الشكل 32** مثال توضيحي عن الارتباطات في wordnet

من الجدير بالذكر أن الـ WordNet تقسم الكلمات إلى أربعة أنواع هي أسماء وأفعال وصفات وأحوال، وتصنف الكلمات تبعاً للأنواع السابقة وتوضع في مجموعات (synsets) حسب المعنى، مع العلم أن العلاقات السابقة الذكر بين المجموعات لا تربط إلا المجموعات التي تحوي الكلمات التي تنتهي إلى نفس النوع، فلأسماء المجموعات الخاصة بها وللأفعال مجموعات خاصة بها.

بعض الإحصائيات عن الـ WordNet

POS	Words	Synsets	Senses
Noun	117097	81426	145104
Verb	11488	13650	24890
Adjective	22141	18877	31302
Adverb	4601	3644	5720
Totals	155327	117597	207016

## جدول 2 إحصائيات عن معجم wordnet

سنوجه اهتمامنا في هذه الدراسة إلى الأسماء والصفات فقط لأنها تحمل معنى أوضح من الأفعال، فال فعل في أغلب الأحيان له معاني أكثر من الاسم، كما أن عدد الأسماء أكثر بكثير من عدد الأفعال.<sup>[8]</sup>

يمكن تحميل معجم الـ WordNet من خلال الرابط التالي : <http://wordnet.princeton.edu/>

domains هو تصنيف لكل synset ضمن ال wordnet بالنسبة لكل ال Wordnet domains -

المتاحة وكم نسبة انتماء هذه ال synset إلى كل domain ،

- يوجد حوالي 170 domain ،

- هناك abstract domains 5 ويتفرع عنها باقي ال domains ،

- من أجل كل domain هناك ملف يحوي كل الكلمات الموجة في ال wordnet مع نسبة انتماء هذه الكلمة إلى

[9].domain هذا ال

#### TOP LEVEL

- > doctrines
- > free\_time
- > applied\_science
- > pure\_science
- > social\_science
- > factotum

## الفصل الخامس

### RDF && DBpedia && Linked Data

---

نعرض في هذا الفصل أنطولوجية dbpedia المستخدمة في فهم الكلمة في حال لم تكن الكلمة موجودة ضمن معجم wordnet، وهنا يتحقق مفهوم المعطيات المترابطة حيث تم ربط الكلمة بمعناها ضمن أنطولوجية dbpedia، سنتعلق لمحة سريعة عن مفهوم rdf حيث يكون محتوى dbpedia على هيئة ثلاثيات .rdf

#### 1 ثلاثيات RDF

- هو اختصار ل (Resource Description Framework) وهي عبارة عن مجموعة من

المواصفات التي وضعها اتحاد الشبكة العنكبوتية (W3C) لتعريف بنية تحتية مرنة خاصة بتنظيم وإدارة

خصائص البيانات التي تسمى (metadata) في الشبكة العنكبوتية؛

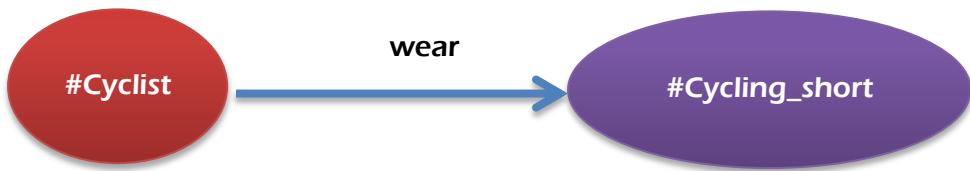
- metadata هي عبارة عن معلومات وصفية عن طبيعة البيانات والوثائق مثل مصدرها، وحجمها و

التنسيق الخاص بها وخصائص أخرى خاصة بالبيانات؛

- إذاً قد تم تصميمها لتوفير إطار يعتمد على لغة XML التي يمكنها أن تقوم بتوحيد عملية تبادل

خصائص البيانات بين التطبيقات المختلفة أو ما يسمى بخصائص المحتويات (metacontent)؛

- ثلاثيات ال RDF تتألف من Subject, Predicate and Object

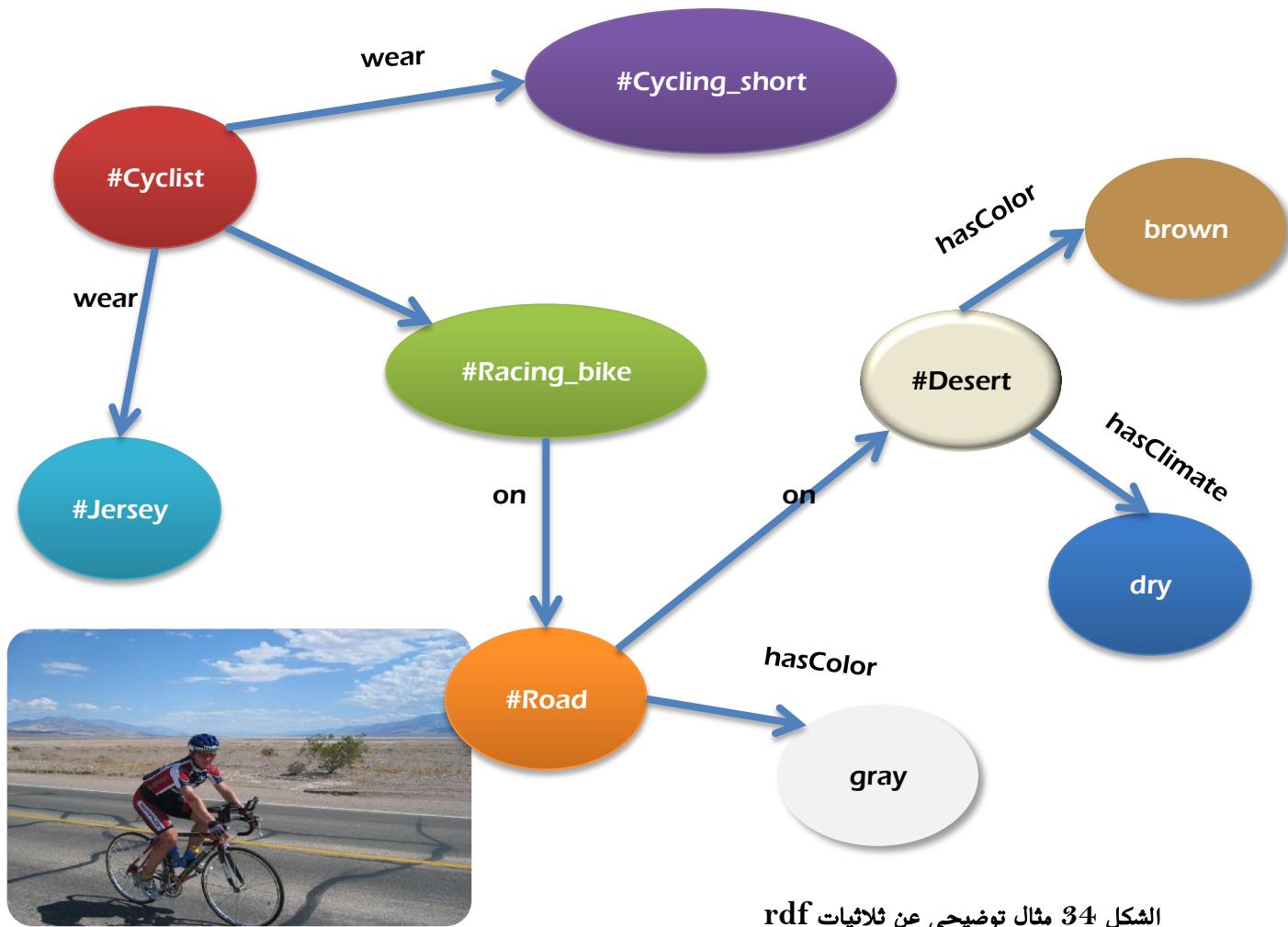


الشكل 33 مثال عن ثلاثية rdf

- ومن الاستخدامات المحتملة لـ RDF محركات البحث، وأنظمة تقييم المحتويات، ومجالات أخرى تهتم

بخصائص البيانات المتبادلة؛

- مثال يوضح ثلاثيات الـ RDF



الشكل 34 مثال توضيحي عن ثلاثيات rdf

- قاعدة معلومات مجانية وضخمة تأخذ البيانات المهيكلة من Wikipedia وتقوم بتحويلها إلى ثلاثيات RDF،
- مقالات Wikipedia تحوي نصوص بطريقة مهيكلة مضمونة ضمن المقالة، يتم استحصال هذه المعلومات المهيكلة ووضعها في بيانات dataset بحيث نستطيع الاستعلام عنها؛
- إحصائيات عن dbpedia في 2011 يقول أنها تصنف حوالي 3.64 مليون موضوع تتوزع هذه الإحصائيات كما في الجدول التالي:

شخص	<b>416,000</b>
مكان	<b>526,000</b>
ألبوم موسيقي	<b>106,000</b>
فيلم	<b>60,000</b>
لعبة فيديو	<b>17,500</b>
منظمة	<b>169,000</b>
مرض	<b>5,400</b>
لغة	<b>97</b>
رابط إلى صور	<b>2,724,000</b>
رابط لصفحات ويب خارجية	<b>6,300,000</b>
رابط إلى بيانات rdf dataset	<b>6,200,000</b>

جدول 3 إحصائيات عن أنطولوجيا DBpedia

1 - Dbpedia إذاً يستخدم ثلاثيات rdf في تمثيل البيانات، إحصائيات أيلول 2011 بأنه يوجد حوالي

665 billion من ثلاثيات ال rdf موزعة بين 385 million ثلاثة مأخذة باللغة الإنكليزية و

million مأخذة بلغات أخرى؟

.SPARQL هي dbpedia ضمن معلومات موجودة عن البيانات المستخدمة في الاستعلامات -

## Linked Data 3

- يتم الربط بين مصادر مختلفة للبيانات مع بعضها البعض والاستفادة منها؛

- تم تطبيق هذه الفكرة في مشروعنا من خلال الربط مع dbpedia،

- استخدمنا لأجل ذلك أداة تقدم لنا خدمة الربط مع dbpedia وهي spotlight،

- في حال لم تكن الكلمة التي نهدف إلى فهم معناها ضمن wordnet يتم البحث عن الكلمة المطلوبة ضمن

dbpedia،

- يتم الاستفادة من العلاقات الموجودة في dbpedia ليتم فهم هذه العلاقات مثل علاقة ال subject، يتم فهم قيم

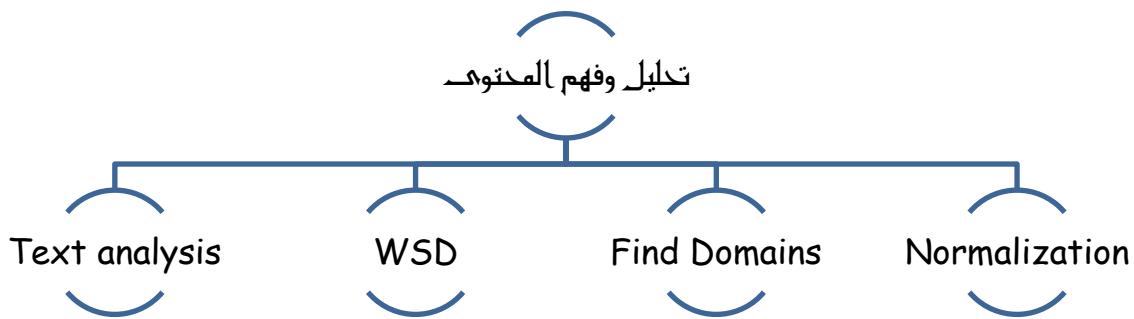
هذه العلاقات من خلال تطبيق عمليات المعالجة الدلالية على هذه العلاقات وفهم الكلمة المطلوبة؛

- بهذه الطريقة تكون قد ربّطنا المعطيات الموجودة لدينا مع شرح عنها ضمن أنطولوجية ال dbpedia.

# الفصل السادس

## دليل وفهم المحتوى

نعرض في هذا الفصل الهدف الأساسي من المشروع والذي يقدمه خدمة تحليل وفهم المحتوى، عملنا على فهم المحتوى من خلال تطبيق خطوات متتالية من عمليات المعالجة الطبيعية ومن ثم الحصول على معاني الكلمات من خلال معجم ال wordnet ، في حال لم تكن الكلمة موجودة يتم البحث عنها ضمن أنطولوجية dbpedia ، بعد ذلك يتم معرفة المجال الذي تنتمي له كل كلمة وزيادة نقاط المجال يقدار انتقاء الكلمة إلى هذا المجال ، وفي النهاية نقوم بعملية normalization للأوزان بحيث نحصل على نسبة انتقاء كل مجال إلى الصفة التي نعمل على فهم وتحليل محتواها.



الشكل 35 النقاط الأساسية في تحليل وفهم المحتوى

## ١ استبدال الضمائر بالكلمات الأصلية التي تعود عليها هذه الضمائر Coreferencing

أحد المشاكل التي واجهتنا أثناء معالجة النصوص، هي ورود اسم شخص مثلاً في بداية الجملة، ثم عند ذكر أفعال أخرى قام بها نفس الشخص فإنه يتم ذكر ضمير عائد على اسم الشخص بدلاً من ذكر اسمه، وهذا يمنعنا من معرفة الفاعل الحقيقي لهذا الفعل مثلاً:

**John** played football, and **he** also played golf.

نلاحظ أن الضمير **he** يعود على الفاعل **John**، لذا نقوم باستبدال الضمير بالفاعل قبل البدء بمعالجة النص وتحويله إلى ثلاثيات فيصبح :

**John** played football, and **John** also played golf.

قمنا بالاستعانة بخدمة **Stanford CoreNLP** الموجودة ضمن مكتبة **Coreference**، لكننا قمنا بتطبيق بعض التحسينات عليها لمعالجه بعض المشاكل مثلاً اذا أدخلنا المثال التالي:

**Damascus** is **the capital of Syria**.

ستكون النتيجة أن جملة **the capital of Syria** كلها تعود على **Damascus**، أي ستتصبح الجملة بعد تطبيق العملية عليها:

**Damascus** is **Damascus**.

وهذه النتيجة غير مجذدة أبداً، لذا قمنا بحصر هذه العملية فقط على الضمائر بكافة أنواعها (**he, she, it, they... they...**)، وإضافة **'s** في حال كان الضمير هو أحد ضمائر الملكية مثل (**his, her ...**).

كمثال على ذلك الجملة التالية التي تملك ضمير الملكية "his" :

John likes all people. So people like **his** personality.

تصبح بعد المعالجة:

John likes all people. So people like **John's** personality.

## 2 إعادة الكلمات إلى أصلها Lemmatization

لزيad من التحسين على النتائج قمنا أيضا بإعادة الكلمات إلى أصلها، حيث قمنا في البداية بتطبيق التجذير stemming على الكلمات لكن استنتجنا أن هذه الطريقة تحوي على العديد من المشاكل، فمثلاً كلمة providers تتحول بتطبيق خوارزمية

provid- وهي أحد خوارزميات التجذير- إلى كلمة porter

حيث نلاحظ أنها قامت بإزالة حرف ال e أيضاً من نهاية الكلمة و بالتالي لم تعد الكلمة صحيحة لغويًا ولا توجد في أي معجم.

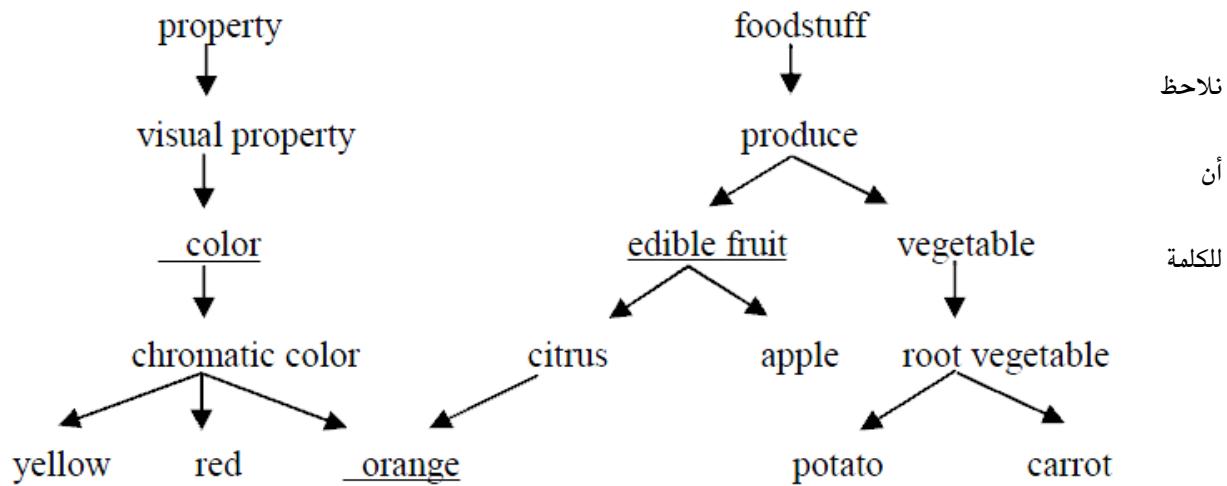
لذلك قمنا باستخدام عملية أخرى تدعى lemmatization حيث تقوم هذه الطريقة بإرجاع الكلمة إلى أصلها لكن بالاستعانة بمعجم مثل WordNet وبالتالي نتائجها تكون صحيحة و موجودة في المعجم.

مثلاً كلمة providers بعد تطبيق العملية عليها تعود إلى الكلمة provider.

## 3 إزالة غموض الكلمات Word Sense Disambiguation

يوجد في جميع اللغات الطبيعية كلمات يمكن أن يكون لها معنى مختلف في سياق مختلف (أكثر من معنى للكلمة الواحدة). فـ غموض الكلمة هي العملية التي تهدف لاكتشاف المعنى الصحيح لكلمة موجودة ضمن جملة بشكل آلي.

على سبيل المثال، لليكن لدينا الشكل التالي:

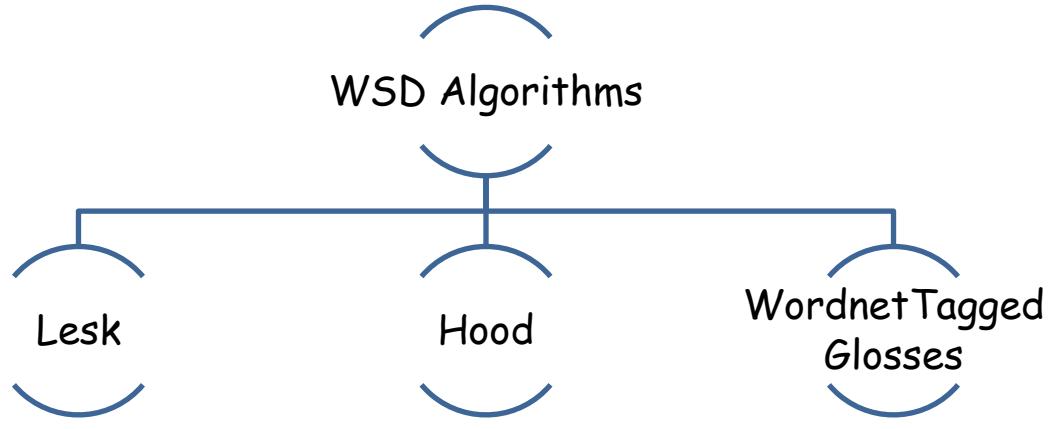


An example of two hypernym trees for the term orange.

الشكل 36 مثال عن غموض كلمة

"orange" شجري hypernym فيمكن أن تأتي بمعنى فاكهة أو لون، وبالتالي فلن نستطيع تحديد المفهوم المقابل لهذه الكلمة، وبالتالي يجب علينا أن نجد المعنى الصحيح للكلمة. هناك العديد من الطرق لتحديد معنى كل كلمة منها ما هو بسيط، ومنها ما هو معقد، والنتائج تتفاوت في الدقة والسرعة.

على سبيل المثال يمكن أن نعتبر أن المعنى الأول لكل كلمة هو المعنى الصحيح، وذلك باعتبار أن معاني الكلمات في ال WordNet مرتبة تنازلياً وذلك حسب تردد استخدام الكلمة بهذه المعاني، فكلمة "dark" تستخدم بمعنى "ظلم" أكثر بكثير من استخدامها بمعنى "مغلق". كما ويمكن استخدام خوارزميات أكثر تعقيد مثل Lesk أو Hood أو غيرها.



الشكل ٣٧ خوارزميات إزالة الغموض

### ٣.١ خوارزمية *Lesk* التقليدية

هذه الخوارزمية مبنية على الافتراضين التاليين :

- ١- عندما تستخدم كلمتين متجلرين في جملة ما، فمن المؤكد إنهما يتكلمان عن نفس الموضوع.
- ٢- إذا كان معنى واحد من كل كلمة يمكن أن يستخدم للتalking عن نفس الموضوع، فلا بد أن يكون تعريف المعجم لكل كلمة يستخدم كلمات مشتركة بينه وبين التعريف الآخر.

وبناءً على هذين الافتراضين فإن خوارزمية تعمل على فك غموض الكلمات الموجودة ضمن جمل قصيرة (Short Phrases)، فمن أجل كل كلمة ضمن الجملة تقوم بإيجاد جميع تعريفات (Glosses) المعاني الخاصة بها (Sences) وذلك بالاستعانة بمعجم لكلمات اللغة الإنجليزية، ومن ثم تقوم بمقارنة كل تعريف من هذه التعريفات مع جميع تعريفات معاني بقية الكلمات في الموجودة في الجملة، إن عملية المقارنة بين كل تعريفين تتم بإيجاد عدد الكلمات المشتركة بينهما، ومنه فإن التعريف المشتركة بأكبر عدد كلمات تكون هي التعريف المناسب لكلمات الجملة.

كمثال على ذلك نأخذ الجملة القصيرة التي تحوي الكلمتين الكلمتين ("pine - cone") فنجد أن هاتين الكلمتين فعلاً يتواجدان معاً وهم يتكلمان عن ("evergreen trees") و بالفعل نجد أنه يوجد في تعريف احدى المعاني لكل كلمة في المعجم الكلمتين : ("evergreen", "tree")

حيث نجد أن للكلمة الأولى pine معنيين هما :

Sense 1: kind of evergreen tree with needle-shaped leaves

Sense 2: waste away through sorrow or illness.

ونجد أن للكلمة الثانية core ثلاثة معاني هي :

Sense 1: solid body which narrows to a point

Sense 2: something of this shape whether solid or hollow

Sense 3: fruit of certain evergreen tree

بمقارنة كل معنى من معاني الكلمة الـ pine مع جميع معاني الكلمة الـ core نجد الكلمتين evergreen tree هما الأكثر تكرارا وبالتالي المعنى الأول هو الأنسب للكلمة pine والمعنى الثالث هو الأنسب للكلمة core.

Pine#1 Ç Cone#1 = 0

Pine#2 Ç Cone#1 = 0

Pine#1 Ç Cone#2 = 1

Pine#2 Ç Cone#2 = 0

Pine#1 Ç Cone#3 = 2

Pine#2 Ç Cone#3 = 0

نلاحظ من خوارزمية Lesk التقليدية أنه على الرغم من بساطتها إلا أنها تعاني من مشكلة أن تعريف المجم قصير جداً، ولا يوجد فيه كلمات كثيرة أو كافية من أجل أن تعمل هذه الخوارزمية بشكل جيد، بالإضافة إلى أنها لا تستفيد من المعاني التي قامت بإسنادها للكلمة خلال عملية إزالة الغموض حيث تقوم بإعادة العمليات من جديد من أجل كل كلمة.

### 3.2 التحسينات على خوارزمية Lesk من خلال Global Disambiguation Using Glosses of Related Synsets

لقد أتت خوارزمية Global Disambiguation بعديد من التحسينات على خوارزمية Lesk التقليدية وذلك ما أدى إلى تحسين واضح في دقة إزالة الغموض وفي السوعة، حيث تعمل هذه الخوارزمية على الاستفادة من الكلمات المحيطة بالكلمة لإزالة غموض الكلمة الحالية، حيث يتم تعريف نافذة ذات طول فردي (الطول الأنساب 3 أي الكلمة الهدف وكلمتين محيطتين واحدة على يمينها وواحدة على يسارها)، ومن التحسينات التي تم إجراءها:

#### 3.2.1 استخدام شبكة دلالية

إن خوارزمية Lesk التقليدية كانت تعتمد على المعاجم للحصول على تعريف معاني كلمات الجملة مثل معجم Oxford، كما ذكرنا أن هذه المعاجم تحوي على تعريف قصيرة غالباً ما تكون غير مجذدة عند استخدامها في خوارزمية Lesk، ولذلك تم اعتماد على الشبكات الدلالية لعنانها بالعلاقات الدلالية التي ستفيدها في الحصول على المزيد من التعريفات المعاني الكلمات والذي بدوره سيؤدي إلى الحصول على المزيد من التقاطعات بين الكلمات ضمن التعاريف.

ومن أشهر الشبكات الدلالية هي Wordnet التي تحوي على العديد من العلاقات التي تربط الكلمات بعضها، وتختلف هذه العلاقات باختلاف موقع الكلمة في الجملة، ومن أشهر هذه العلاقات علاقتي Hyponym و Hypernym والتي هي

مسؤولة عن ترابط الكلمات بهيكلية is-a والتي ستفيدنا بالحصول على المزيد من التعاريف للكلمة الواحدة من خلال تعاريف أباءها

Noun	Verb	Adjective
Hypernym	Hypernym	Attribute
Hyponym	Troponym	Also see
Holonym	Also see	Similar to
Meronym		Pertainym of
Attribute		

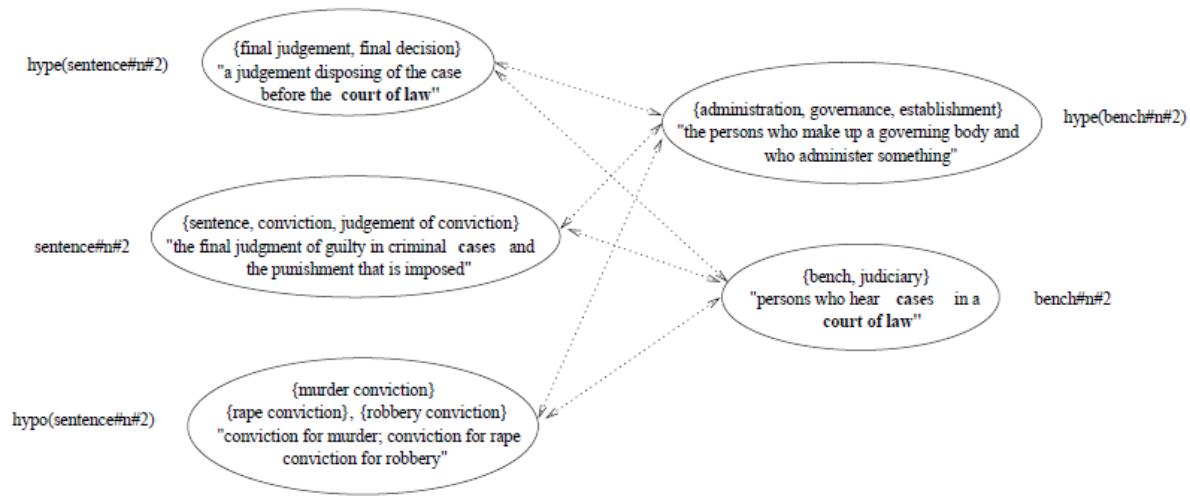
وأبنائهما في الهيكلية. [١٧]

والجدول التالي يوضح أهم العلاقات التي من الممكن استخدامها مصنفة حسب موقع الكلمة من الكلام:

#### أساليب اختيار التعاريف الواجب مقارنتها

هناك عدة أساليب لاختيار التعاريف التي حصلنا عليها من الشبكة الدلالية التي يجب مقارنتها، فإذا أخذنا تعريفين لكلمتين وأوجدنا جميع التعاريف المرتبطة بهما بالعلاقات التي ذكرناها، فإنه من الممكن أن نقوم بمقارنة كل تعريف تابع لكلمة أو لإحدى علاقتها مع التعاريف المرتبطة بالكلمة الأخرى ويدعى هذه الأسلوب Heterogeneous.

إذا أخذنا الكلمتين Hyponym و Bench وبأخذ العلاقات بينهما Hyponym و Sentence المسمى hypo فإنه سيتم المقارنة بين الكلمتين وفق الشكل التالي:



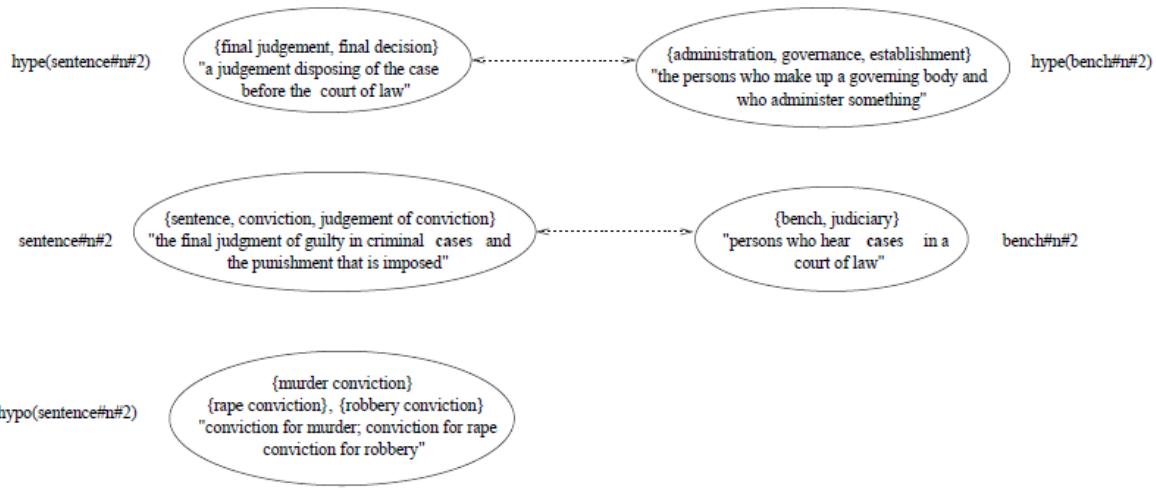
**الشكل 38 الدمج بين المجموعات**

إن الكتابة sentence#n#2 ترمز إلى أن الكلمة sentence هي من النوع Noun وأن هذا هو المعنى الثاني لها.

كما بالإمكان اتباع أسلوب آخر في المقارنة حيث يتم فيه مقارنة التعريف التي تنتهي إلى نفس الفئة فمثلا التعريف التي تم إيجادها

بالعلاقة hype في كلا الكلمتين يتم مقارنتها مع بعضها فقط، والتعريف التي تم إيجادها بالعلاقة hypo تتم مقارنتها سويةً

فقط ويسمى هذا الأسلوب Homogeneous والشكل التالي يوضح أسلوب المقارنة هذا أكثر بين الكلمتين في المثال السابق:



**الشكل 39** أسلوب المقارنة بين المجموعات

نلاحظ من الشكل السابق أنه لم نجد للكلمة **bench** أي رابط مع كلمات أخرى بعلاقة **hypo** ولذلك فإنه لن يتم المقارنة عبر هذه العلاقة.

من الأسلوبين السابقين نلاحظ أننا في الأسلوب الأول **Heterogeneous** سنحصل على تقاطعات بين الكلمتين أكثر من الأسلوب الثاني **Homogeneous** ولكن في الأسلوب الثاني سنحصل على سرعة أكثر نظراً لقلة عدد المقارنات. [18]

#### 3.2.2 ايجاد التقاطعات بين تعريفين

إن طريقة **Lesk** التقليدية كانت تعتمد على ايجاد عدد الكلمات المشتركة بين التعريفين، في هذه الطريقة أوجدنا أسلوب جديد لحساب التقاطعات بالاعتماد على أطول السلسل المشتركة بين التعريفين ومن ثم توزينها وهذا الأسلوب يعتبر أكثر مجدياً من الطريقة التقليدية لإبرازه أهمية التقاطعات.

فلو كان لدينا التعريفين التاليين:

Sentence 1: you are the boy

Sentence 2: you are the boy I like

بالطريقة التقليدية نحصل على 4 كلمات متشابهة بين التعريفين وبالتالي سيكون الرصيد (score) هو 4 وهو نفس الرصيد الذي سنحصل عليه فيما اذا كانت الكلمات متباude عن بعدها ضمن كل تعريف، في حين أننا لو أخذنا المشترك هو أطول سلسلة كلمات مشتركة بين الجملتين فسنحصل على جملة واحدة مكونة من 4 كلمات، وفي حين كانت الكلمات متباude سنحصل على 4 كلمات، وهنا نجد أن طريقة السلسلة المشتركة الأطول تأخذ بعين الاعتبار ورود الكلمات المشتركة مع بعضها بحيث تعطيها أهمية أكبر عن الكلمات فيما لو كانت متباude.

ولتحقيق هذه الأهمية لا بد من تربيع الرصيد الذي نحصل عليه من كل تشابه ، ففي المثال السابق يصبح الرصيد  $4 * 4 = 16$  هو رصيد الجملة، نلاحظ أنه أكبر من الرصيد الذي نحصل عليه فيما لو كان لدينا تشابهين أحدهما هو كلمة واحدة والآخر هو 3 كلمات حيث سيكون الرصيد هو  $(1 * 3) + (1 * 3) = 10$  ، وأيضاً هذا الرصيد أكبر فيما لو كانت الكلمات الأربع متباude عن بعضها حيث حيث  $(1 * 1) + (1 * 1) + (1 * 1) + (1 * 1) = 4$ . وبالتالي نجد كيف أن هذا الأسلوب فعلاً قد أعطى أهمية أكبر لورود الكلمات المتشابهة خلف بعضها عن الكلمات المتشابهة المتفرقة.

وقد تم استنتاج هذه الطريقة في الحساب من قانون الجبر التالي :

$$(a_0 + a_1 + \dots + a_n)^2 > a_0^2 + a_1^2 + \dots + a_n^2$$

### 3.2.3 استراتيجية عمل خوارزمية Global Disambiguation

إحدى أبرز السينات في خوارزمية Lesk التقليدية كما ذكرنا سابقاً هو أنها لاستفادة من المعاني التي قامت بإسنادها للكلمة خلال عملية إزالة الغموض حيث تقوم بإعادة العمليات من جديد من أجل كل كلمة، وهذا ما تم معالجتها أيضاً في هذه الخوارزمية حيث يتم تحديد نافذة من الكلمات المراد معالجتها سوية واستخراج المعاني الأنسب لجميع كلمات النافذة سوية، أي أن كل كلمة في النافذة تساعده في إزالة غموض الكلمات الأخرى الموجودة ضمنها، وإن الحجم الأفضل للنافذة كما ذكرنا سابقاً هو 3 وذلك بسبب زيادة تعقيد العمليات الحسابية كلما زاد حجم النافذة أكثر.

بعد تحديد حجم النافذة نقوم بإيجاد جميع المعاني لكل كلمة من كلمات النافذة، ومن ثم إيجاد جميع التراكيب المكونة من معاني الكلمات مع بعضها، ومن ثم نقوم بإجراء مجموعة من عمليات المقارنة على كل تركيبة من المعاني وحساب رصيدها. في النهاية وبعد إيجاد رصيد كل التراكيب المكونة نقوم باختيار أكبر رصيد موجود لتكون التركيبة الموافقة له هي التي تحمل المعاني الأنسب للكلمات الموجودة ضمن النافذة.

### حساب رصيد التركيبة

1. نقوم بإيجاد جميع الأزواج المكونة من الكلمات الموجودة ضمن كل تركيبة
2. من أجل كل زوج من الكلمات
  - a. نقوم بإيجاد جميع العلاقات المرتبطة بكل كلمة في الزوج
  - b. حسب أسلوب المقارنة المحدد (Heterogeneous, Homogeneous) نقوم بتحديد كيفية المقارنة للتعريف التي تم ايجادها لكلمات الزوج والكلمات التي أوجدناها بالعلاقات
  - c. نقوم بإيجاد التقاطعات بين كل تعريفين اعتماداً على الأسلوب الذي ذكرناه بالتحسين (أطول سلسلة مشتركة) لنحصل على رصيد نقوم بجمعه إلى رصيد التركيبة التي ينتمي إليها هذا الزوج.

ملاحظة: قد تكون المجموعة (synset) التي تحوي الكلمة الواحدة مرتبطة بعلاقة Hypernym (أو أي علاقة أخرى) مع عدة مجموعات أخرى (أي يكون لدينا عدة تعاريف)، عندئذ نقوم بجمع كل هذه التعريفات للمجموعات المرتبطة بنفس العلاقة بتعريف واحد يكون هو التعريف المثل لارتباط الكلمة بهذه العلاقة مع بقية الكلمات.

مثال على تطبيق الخوارزمية :

لتكن لدينا نافذة ذات طول 3 ملزمة من الكلمات التالية : sentence, offender, bench

بإيجاد معاني كل كلمة من الكلمات الثلاث نحصل على:

gloss(sentence#n#1) = a string of words satisfying the grammatical rules of a language

gloss(sentence#n#2) = the final judgment of guilty in criminal cases and the punishment that is imposed

gloss(bench#n#1) = a long seat for more than one person

gloss(bench#n#2) = persons who hear cases in a court of law

gloss(offender#n#1) = a person who transgresses law

بإيجاد جميع التراكيب الممكنة من هذه المعاني نحصل على المجموعة التالية:

Combination 1: sentence#n#1 – bench#n#1 – offender#n#1

Combination 2: sentence#n#1 – bench#n#2 – offender#n#1

Combination 3: sentence#n#2 – bench#n#1 – offender#n#1

Combination 4: sentence#n#2 – bench#n#2 – offender#n#1

الآن نقوم بأخذ كل تركيبة وإيجاد جميع الأزواج الممكنة وإجراء عملية المقارنة على كل زوج وحساب رصيد كل تركيبة، فمثلاً إذا

أخذنا التركيبة الأخيرة رقم 4 فإننا سنحصل على الأزواج التالية الموضحة بالجدول الآتي:

First Gloss	Second Gloss	Overlap String	Normalized Score
hype(sentence#n#2)	bench#n#2	court of law, case	10
sentence#n#2	bench#n#2	cases	1
hype(sentence#n#2)	offender#n#1	law	1
sentence#n#2	hypo(offender#n#1)	criminal	1
hype(bench#n#2)	hype(offender#n#1)	person	1
hype(bench#n#2)	offender#n#1	person	1
hype(bench#n#2)	hypo(offender#n#1)	person	1
bench#n#2	hype(offender#n#1)	person	1
bench#n#2	offender#n#1	person, law	2
bench#n#2	hypo(offender#n#1)	person	1
Total score for sentence#n#2 – bench#n#2 – offender#n#1			20

#### الشكل 40 حساب رصيد كل تركيبة

نلاحظ من الجدول أن الأسلوب المتبع في مقارنة الأزواج هو Heterogeneous.

وبالتالي نجد من الجدول أن رصيد التركيبة رقم 4 هو 20، وبالمثل نقوم بإيجاد رصيد كل التركيبات الأخرى و اختيار الرصيد

الأكبر لتكون التركيبة الموافقة هي الممثلة لكلمات النافذة.

#### تعقيد الخوارزمية

يحسب تعقيد الخوارزمية بالعلاقة التالية:

$$S^2 * \frac{N*(N-1)}{2}$$

حيث N : هو عدد الكلمات ضمن النافذة

S : هو وسطي عدد معاني كل الكلمة

### 3.3 خوارزمية Hood

هي طريقة لتصنيف المستندات بشكل آلي مبنية على فك غموض الكلمات automatic text classification

حيث يتم فك غموض مجموعة من الكلمات معاً بنفس time based on word sense disambiguation

الوقت، وتبدل كل كلمة بالمعنى الصحيح لها، و هنا نجد اختلاف هذا الخوارزمية عن خوارزمية primitive

حيث أن هذه Weighted Overlapping و خوارزمية adaptive lesk و خوارزمية lesk

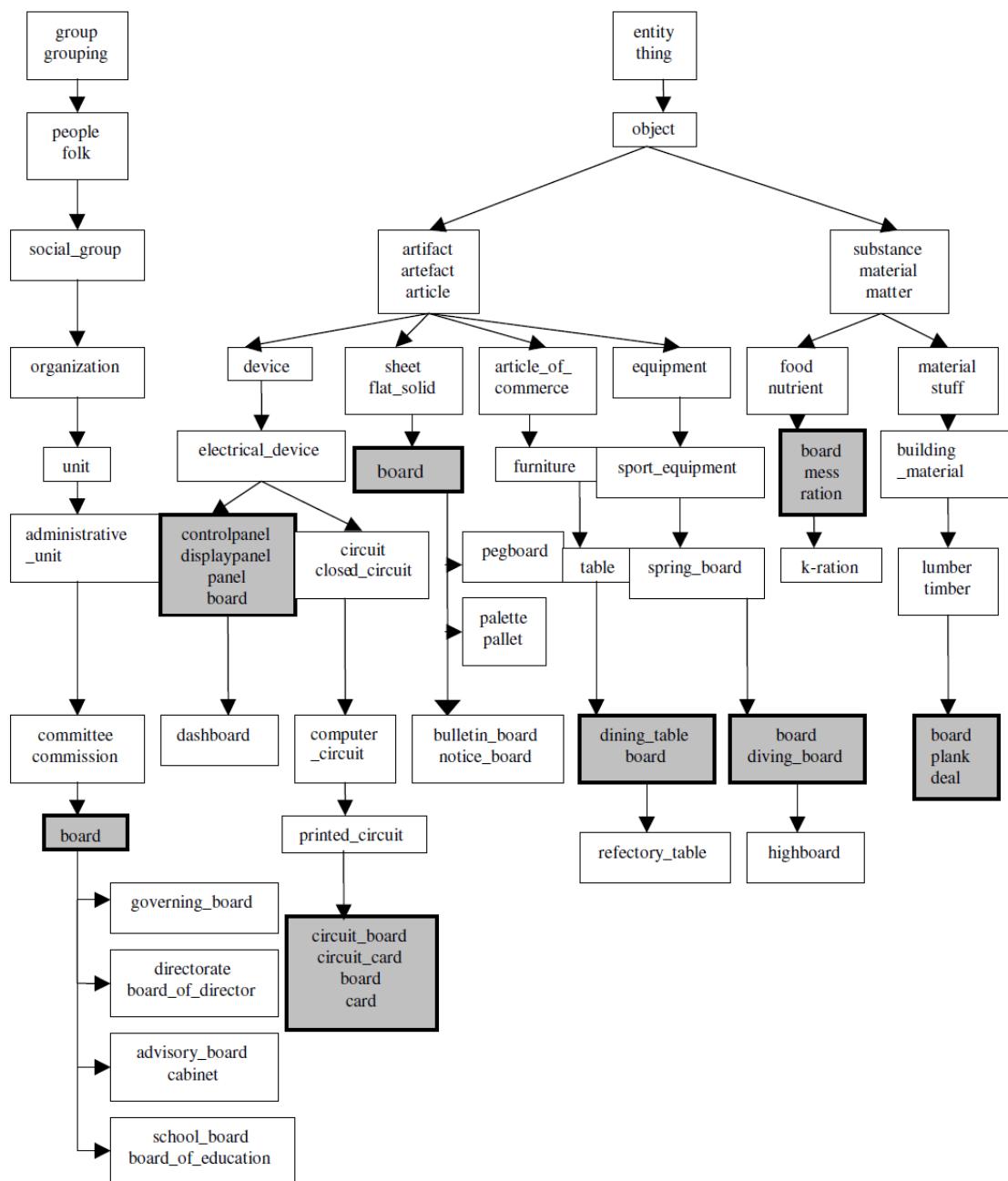
الخوارزميات تقوم بفك غموض الكلمات بشكل يدوي أو فردي manual (تأخذ الكلمات المراد فك غموضها كلمة كلمة

حيث يتم فك غموض أول الكلمة ثم التالية ) أما خوارزمية Hood فيتم فك الغموض لمجموعة الكلام بنفس الوقت ،

ويتم اعتبار أقرب سلف ancestor للمعنى الصحيح صف من صنوف تصنيف المستند ، مع العلم أن أغلب الأبحاث في

word sense disambiguation تركز على استرجاع و الاستعلام عن المعلومات وليس على تصنيف المستندات

. (النصوص)



الشكل ٤١ مثال عن إيجاد جذر hood

The IS-A hierarchy for eight different senses of the noun "board" with the root hood of each synset of the word "board"

### فک غموض الكلمة وتصنيف المستندات word sense disambiguation and text classification 3.3.1

خوارزمية Hood مبنية على فكرة أن مجموعة من الكلمات المتواجدة معا سوف تحدد المعنى المناسب لكل كلمة منها ،

على الرغم من أن كل الكلمة لها أكثر من معنى .

مثال : الكلمات التالية hit and glove ، bat.base كل الكلمة منها لها أكثر من معنى و لكن عندما تكون معا

فأن الموضوع هو لعبة البيسبول .

وبالتالي لاستغلال هذه الفكرة بشكل آلي، يجب تعريف مجموعة من الصنوف (classes) تمثل مختلف المعاني

للكلمات ، ثم نأخذ عدد يقون بعد الكلمات التي لها هذا المعنى و بالتالي المعنى الصحيح للكلمة يتحدد بالصف الذي له

أكبر قيمة عداد و أقرب سلف لهذا المعنى يكون تصنيف للمستند.

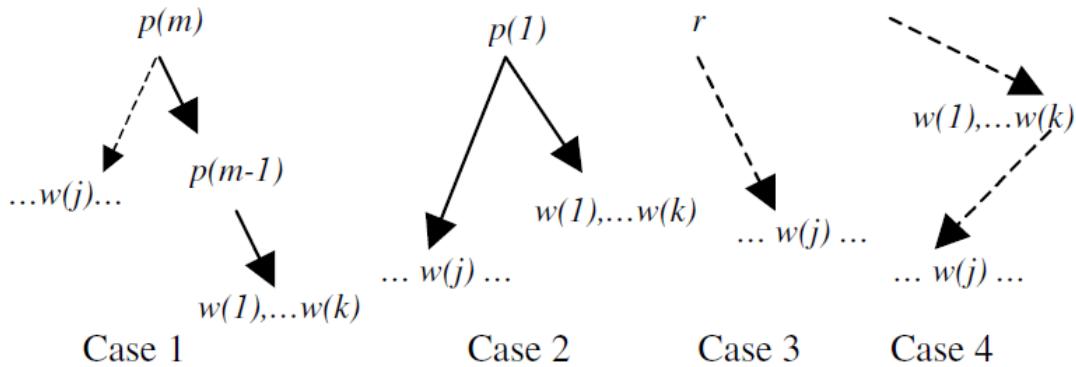
### (Hood Construction) Hood إنشاء 3.3.2

يعرف الـ Hood لمجموعة مترافات synset بأنه أكبر بيان جزئي متصل يحوي synset ويحوي أبناء من

أسلاف للـ synset ، ولا يحوي أية مجموعة مترافات أخرى تملك ابن يملك كلمة موجودة في synset ، ويمثل

الـ Hood Root بالجذر Hood .

### حالات تحديد الجذر للـ Hood



الشكل 42 حالات تحديد الجذر في خوارزمية hood

الحالة الأولى : أن يكون أحد أسلاف الـ synset يملك ابن يحتوي كلمة من كلمات الـ synset وبالتالي الجذر هو

السلف السابق لهذا السلف (case 1).

الحالة الثانية : أن تكون الـ synset هي نفسها الجذر لأن السلف المباشر (الأب) لها يملك ابن يحتوي كلمة من كلمات

الـ synset . (case 2)

الحالة الثالثة : أن يكون جذر الهيكلية للـ WordNet هو الجذر اذا لم يوجد سلف يملك ابن يحتوي كلمة من كلمات

الـ synset مع عدم الإخلال بشرط تعريف الـ Hood . (case 3)

الحالة الرابعة : اذا كانت الـ synset نفسها تملك ابن يحتوي كلمة من كلمات الـ synset وبالتالي لا يوجد

لهذه الـ synset [19]. (case 4)

– من أجل كل الكلمة خامضة:

- من أجل كل sense لهذه الكلمة يتمأخذ جذر hood المقابل:
  - نمرعلى جميع الـ synsets لجميع الكلمات المراد فك غموضها، ونأخذ علاقات الـ hypernyms لهذا الـ synsets ، وعند المرور Hood root يتم زيادة العدد لهذا الـ Hood Root قيمة واحد وبالنهاية يكون المعنى الصحيح للكلمة هو المعنى المتواجد ضمن synset وجذر hood المقابل له هو الجذر ذو العدد الأعلى.

## 3.4 إزالة الغموض بالاستعانة بـ WordNet Semantically Tagged glosses

### 3.4.1 لحة عن مشروع WordNet Semantically Tagged glosses

تعتمد هذه الطريقة على مشروع WordNet Semantically Tagged glosses والذي تم اطلاقه من قبل الجامعة المنتجة لأنطولوجيا WordNet، الهدف من هذا المشروع هو إزالة غموض جميع كلمات التعريف glosses الخاصة بـ SynSet، اي انه يقوم بربط جميع كلمات التعريف بالمعنى الصحيح المقابل لها بالاستفادة من الكلمات المحيطة.

فلنأخذ مثلاً كلمة orange ونأخذ تعريف أحد المعاني الخاصة بهذه الكلمة:

Round yellow to orange fruit of any of several citrus trees

يهتم المشروع بإزالة غموض جميع الكلمات المطللة بشكل مسبق، ويقوم بربطها بالمعنى الصحيح لها ضمن WordNet.

نقوم بتسمية المعاني الناتجة عن إزالة غموض تعريف معين بـ المعاني المرتبطة بذلك المعنى، اي ان المعاني المرتبطة بـ معنى round, fruit, citrus هنا هي المعاني المناسبة لكل من orange

وبالتالي يقدم لنا المشروع corpus ضخم جداً من النصوص التي تمت عملية إزالة غموض كلماتها بشكل مسبق، ويمكننا الاستفادة منها بكثير من المجالات منها إزالة غموض نصوص أخرى.<sup>[14]</sup>

### 3.5 الخوارزمية المتبعة لإزالة الغموض :

تقوم فكرة الخوارزمية على أنه عادةً ما تأتي المعاني المرتبطة بمعنى معين ضمن الكلمات المحيطة بهذا المعنى، وهذا يساعد على معرفة المعنى الحقيقي للكلمة من بين جميع المعاني المحتملة

لفرض مثلاً أنه لدينا الجملة التالية ونحاول معرفة المعنى الصحيح لكلمة pen ضمن هذه الجملة:

The **pen** is taking care of the small swans.

• نبحث عن المعاني المحتملة لكلمة pen ضمن WordNet فنجد أنها يمكن أن تأخذ المعاني التالية:

1. a writing implement with a point from which ink flows
2. an enclosure for confining livestock
3. a portable enclosure in which babies may be left to play
4. a correctional institution for those convicted of major crimes
5. female swan

• نوجد الآن المعاني المرتبطة بكل معنى من هذه المعاني بالاستعانة WordNet Semantically Tagged :glosses

1. a writing implement with a point from which ink flows
  - a. point
  - b. ink
  - c. writing implement
2. an enclosure for confining livestock

- a. enclosure
  - b. livestock, stock, farm animal
3. a portable enclosure in which babies may be left to play
- a. portable
  - b. enclosure
  - c. baby, babe, infant
4. a correctional institution for those convicted of major crimes
- a. major
  - b. crime, offense, criminal offense, criminal offence, offence, law-breaking
  - c. correctional institution
5. female swan
- a. female
  - b. swan

• نقارن المعاني المرتبطة بكل معنى من المعاني الخمسة لكلمة pen مع جميع المعاني المحتملة للكلمات المحيطة بكلمة pen

في الجملة المدخلة و نجمع 1 لل score الخاص بأحد المعاني الخمسة في حال حصل تطابق

• نلاحظ ان المعنى الخامس يأخذ ال  $score = 1$  لتطابق المعنى swan من المعاني المرتبطة به مع احد معاني الكلمة

swan الموجودة في الجملة المدخلة، و المعاني الاربعة الأخرى تأخذ ال  $score = 0$  لذلك يكون المعنى الحقيقي للكلمة

هو المعنى الخامس

• في حال تساوى اكثرا من معنى بال score وكان هذا اعلى score نأخذ المعنى الاكثر استخداما في اللغة (يمكننا الحصول

عليه من WordNet) من بين هذين المعنيين. [15]

### 3.5.1 ايجابيات وسلبيات هذه الخوارزمية

تتميز هذه الخوارزمية بسرعتها العالية كونها تعتمد على علاقات محسوبة بشكل مسبق وتقوم فقط بعمليات مقارنة بسيطة، الا انها لا تقدم دقة عالية في النتائج، مما يجعلها مناسبة للتطبيقات التي تحتاج لسرعة عالية و لا تطلب دقة كبيرة في النتائج.

### 4 مقارنة بين الخوارزميات الأربع المستخدمة في إزالة الغموض

الفعالية	السرعة	اسم الخوارزمية
		خوارزمية المعنى الأكثر شهرة
		Wordnet tagged glosses
		Hood
		خوارزمية Lesk

جدول 4 مقارنات بين الخوارزميات الاربعة في إزالة الغموض

## 4.1 طريقة تحليل وفهم المحتوى اعتماداً على معالجة اللغات الطبيعية

### 4.1.1 قراءة صفحة الإنترنت المطلوبة وتحويلها إلى ملف نصي

استخلاص النص من صفحة الويب المطلوبة سواء كان .....text, doc, pdf, docx, html.....

### 4.1.2 تجزئة الكلمات

يتم تجزئة الكلمات التي تحوي على أحرف كبيرة في منتصفها أو التي تحوي المحرف "—" فذلك يعني أنها كلمة جاءت كتركيب أو أنها اسم علم؛

### 4.1.3 إزالة الحروف الغريبة

تتم عملية إزالة المحارف الغريبة من النص مثل المحارف التي لا تنتمي إلى الأبجدية الإنجليزية أو الأرقام لأن لا معنى لها ضمن معالجتنا للنصوص؛

### 4.1.4 تقسيم النص إلى جمل

يتم تقسيم النص إلى جمل حسب علامات الترقيم حيث يتم تمييز جملة عن أخرى من خلال المحرف ". ." وتمييز الكلمات ضمن الجمل ومعرفة موقع الكلمة ضمن الجملة هل هي فعل أم اسم أم صفة أم ظرف؛

### 4.1.5 المعالجة الدلالية وعملية الفلترة

● من أجل كل جملة من الجمل التي تم الحصول عليها من المرحلة السابقة

○ من أجل كل كلمة token ضمن الجملة؛

■ يتم استبدال الضمائر بالكلمات الأصلية التي تعود عليها هذه الضمائر Coreferencing وذلك

ضمن خيار يتم تفعيله أو إلغاء تفعيله من قبل مستخدم النظام كون هذه العملية تستغرق وقتاً؛

■ التحقق من موقع الكلمة ضمن الكلام هل هو اسم أو صفة؛

■ التتحقق من أن هذه الكلمة هي كلمة تحمل معنى أي أنها لا تنتمي إلى مجموعة ال stopwords

- التحقق من أن طول الكلمة يتجاوز المحرفين كون أن أي كلمة تتتألف من محرفين فقط هي كلمة لا تحمل معنى ويجب الاستغناء عنها؛
- الحصول على أصل الكلمة من خلال عملية ال Lemmatization وذلك بهدف الحصول على كلمة موجودة ضمن معجم wordnet والاستغناء بذلك عن عملية ال stemming؛
- محاولة إزالة غموض هذه الكلمة والحصول على المعنى الحقيقي لها حسب سياق الجملة الواردة ضمنها؛
- إيجاد المجالات التي تنتمي لها الكلمة domain من خلال الاستعانة ب wordnet
- ومن ثم الحصول على وزن هذه الكلمة ضمن كل domain والانتقال إلى شجرة domains المجالات وزيادة مجموعة كل domain بوزن الكلمة ضمن هذا ال domain وزيادة مجموعة المجالات الآباء التي يتبع لها ال domain الحالي؛
- في حال لم تكن الكلمة موجودة ضمن معجم ال wordnet نحاول البحث عنه ضمن أنطولوجية dbpedia والاستفادة من علاقة ضمن هذه الأنطولوجية والتي تشرح معنى هذه الكلمة حيث يتم إعادة تطبيق نفس الخطوات من أجل شرح معنى هذه الكلمة ضمن dbpedia.
- نقوم بعملية normalization للأوزان التي حصلنا عليها بعد تطبيق المراحل السابقة ضمن شجرة wordnet domains tree حيث يتم إعطاء ال domain الحاصل على أعلى وزن القيمة 1 وإعادة توزين ال domains tree المتبقية اعتماداً على نسبة مئوية من ال domain الأعلى وزن.
- بعد ذلك يتم تمرير شجرة domains tree إلى مجموعة القواعد التي تم وضعها من قبل خبراء وبناء على رغبة مستخدمي هذا المنتج ليتم معالجة تلك القواعد اعتماداً على الشجرة الممررة ومعرفة هل سيتم حجب صفحة الويب المطلوبة أم لا.

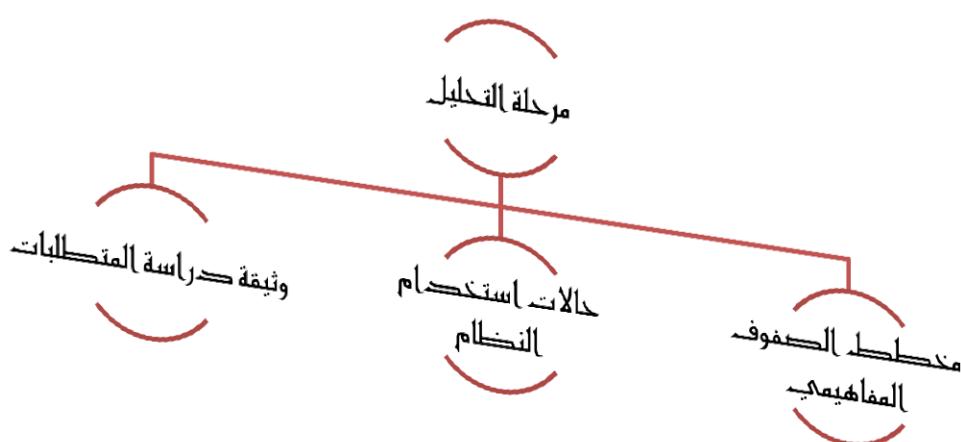
## 4.2 طريقة فهم المحتوى اعتماداً على ثلاثيات RDF

- 1- استخلاص النص من صفحة الويب المطلوبة سواء كان .....text, doc, pdf, docx, html.....
- 2- تجزئة الكلمات التي تحوي على أحرف كبيرة في منتصفها أو التي تحوي المحرف "—" ذلك يعني أنها كلمة جاءت كتركيب أو أنها اسم علم؛
- 3- إزالة المحارف الغريبة من النص مثل المحارف التي لا تنتمي إلى الأبجدية الإنكليزية أو الأرقام لأن لا معنى لها ضمن معالجتنا للنصوص؛
- 4- تقسيم النص إلى جمل حسب علامات الترقيم حيث يتم تمييز جملة عن أخرى من خلال المحرف ". " وتمييز الكلمات ضمن الجمل ومعرفة موقع الكلمة ضمن الجملة هل هي فعل أم اسم أم صفة أم ظرف؛
- 5- يتم تحويل النص إلى ثلاثيات subject, predicate and object؛
- 6- يتم ربط subject, predicate مع أنطولوجيات معروفة مثل dbpedia و URIs هي صلة الوصل ما بين هذه الثلاثيات والأنطولوجيا؛
- 7- يتم فهم محتوى هذه الثلاثيات من خلال العلاقات الموجودة بالأنتولوجية ومحاولة معرفة الـ wordnet domain الخاص بها من خلال هذه العلاقات ومن ثم الحصول على وزن هذه الكلمة ضمن كل domain والانتقال إلى شجرة المجالات وزيادة مجموع كل domain بوزن الكلمة ضمن هذا الـ domain وزيادة مجموع المجالات الآباء التي يتبع لها الـ domain الحالي؛
- 8- نقوم بعملية normalization للأوسمان التي حصلنا عليها بعد تطبيق المراحل السابقة ضمن شجرة wordnet domains tree حيث يتم إعطاء الـ domain الحاصل على أعلى وزن القيمة 1 وإعادة توزين الـ domains المتبقية اعتماداً على نسبة مئوية من الـ domain الأعلى وزن؛

9- بعد ذلك يتم تمرير شجرة tree domains إلى مجموعة القواعد التي تم وضعها من قبل خبراء وبناء على رغبة مستخدمي هذا المنتج ليتم معالجة تلك القواعد اعتماداً على الشجرة المتررة ومعرفة هل سيتم حجب صفحة الويب المطلوبة أم لا.

## الباب الثالث

### مرحلة التحليل



الشكل 43 النقاط الأساسية ضمن مرحلة التحليل



## الفصل السابع

### وثيقة رأسة متطلبات النظام

وهي عبارة عن وثيقة تكتب من قبل الشركة المطورة للمنتج، توضح من خلالها مميزات المنتج الذي تقوم ببنائه، وتعتبر وثيقة تفاهم بين الشركة المطورة للمنتج والشركات التي ترغب بالحصول عليه.

من خلال هذه الوثيقة يتم تحديد متطلبات المشروع ويتم الاستفادة من خبرات المهندسين من أجل تزويد المشروع بأفضل الحلول من أجل تحقيق هذه المتطلبات.

#### 1 تعريف المشكلة المدروسة

تزايد كبير ومستمر في صفحات الإنترنت والمحتوى الضخم الموجود على الشبكة العنكبوتية، جعل من غير الممكن القيام بعمليات تحليل وفهم مسبق للمحتوى، عمليات المعالجة التي تطبق على المحتوى من فلترة وتصنيف وغير ذلك من غير العمليي والمنطقى أن يتم تطبيقها بشكل مسبق.

يمكن القيام بهذه العمليات بشكل يدوى لكن الموضوع ليس بهذه السهولة بسبب النمو الكبير المتزايد، ستكون النتيجة غير منطقية ولن تؤدي إلى النتيجة المطلوبة، إذا لابد من فهم دلالي آلي آني للمحتوى وأخذ النتيجة بشكل لحظي.

## 2 الغاية من بناء هذا المنتج

إذا سيكون الحل هو بناء منتج يعمل على الفهم الدلالي الآلي للمحتوى، هذه الخدمة التي سيتم تقديمها من خلال المشروع سيتم استغلالها في عدة نواحي من عمليات معالجة المحتوى، سواء على مستوى الفلترة، أو على مستوى التصنيف، أو على مستوى محركات البحث أيضا.

## 3 تحديد الجهة المهمة بالمشروع

1- شركة ترغب بحجب محتوى ما عن موظفيها أو مستخدم الحواسيب ضمن الشركة وضمن الشبكة الخاصة بالشركة، من

خلال Proxy server يتم من خلال الحجب وعدم الحجب بناء على محتوى الصفحة وبناء على القواعد الموضوعية

من قبل مديرية هذه الشركة "أي للاستخدام الجماعي"؛

2- شخص يرغب بحجب محتوى ما عن أولاده أو مدرس عن طلابه وغير ذلك من الأشخاص الذين يرغبون بالحجب، من

خلال إضافة يتم إضافتها على المتصفح chrome extention "الاستخدامات الفردية"؛

3- جهة ترغب بتصنيف المحتوى الموجود لديها وتوزيعه على مجالات محددة وذلك من خلال تطبيق موبايل يعمل على

تصنيف الأخبار الواردة حسب المجال الذي تدرج تحته.

## 4 أجزاء المشروع الأساسية

يتتألف المشروع كمنتج نهائي من خدمة ويب web service تعمل على الفهم الدلالي لمحتوى، تم الاستفادة من هذه

الخدمة من خلال ثلاثة تطبيقات:

1- فلترة صفحات الإنترنت من خلال هذا المخدم الذي يخدم كل الموجودين على الشبكة ضمن شركة مثلا

Proxy server

- 2- إضافة على المتصفح تعمل على فلترة صفحات الإنترنت المطلوبة من خلال المتصفح Chrome extention،
- 3- تطبيق موبايل يهدف إلى تصنیف الأخبار الواردة من قناة ما.



## الفصل الثامن

### حالات استخدام النظام

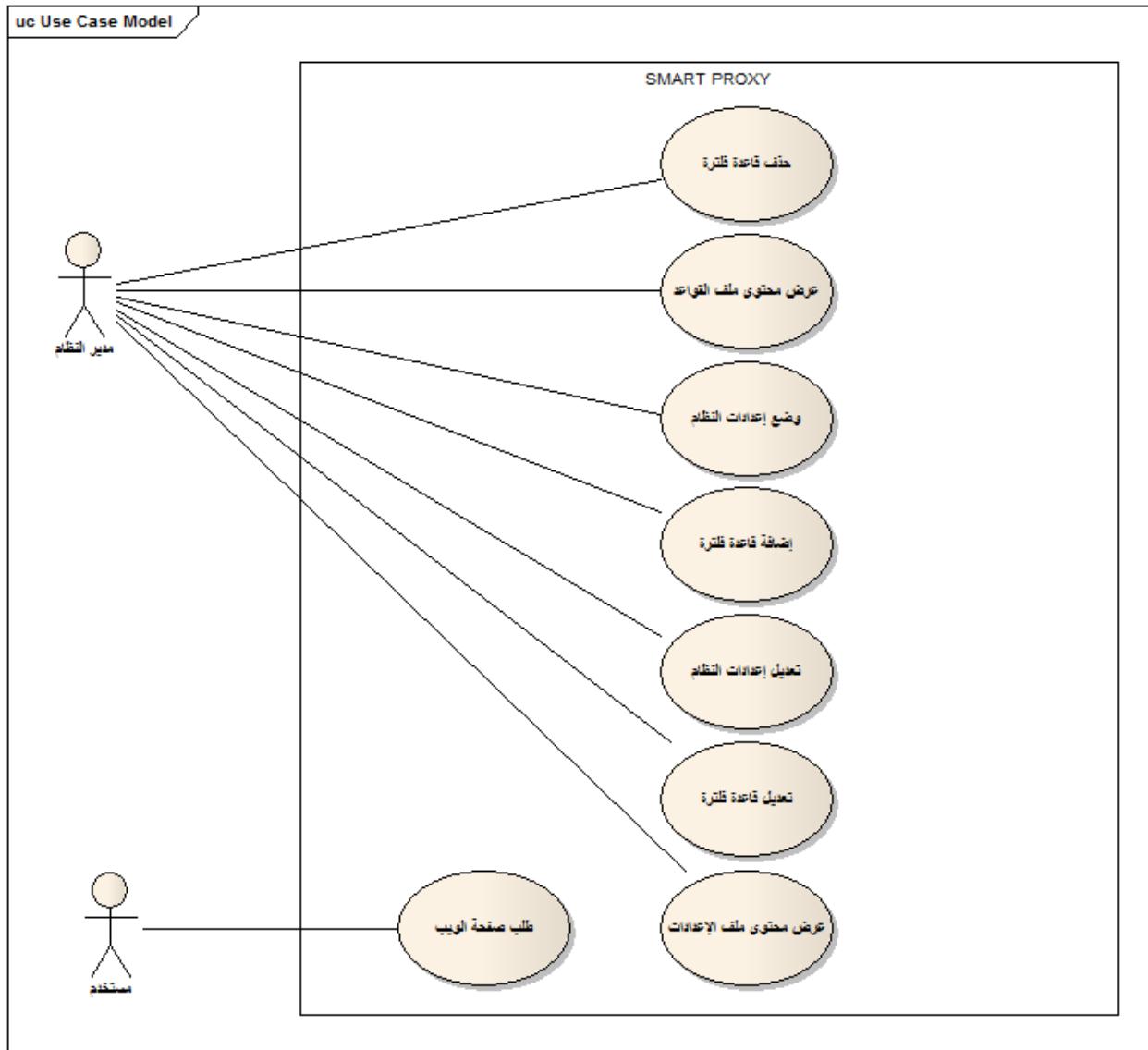
1 مقدمة

مخطط حالات الاستخدام يظهر نقاط التفاعل بين النظام وبين مستخدمي النظام، وهي تعطي فكرة أولية عن آلية عمل النظام من وجهة نظر المستخدم، فلا يظهر من النظام إلا العمليات التي يتفاعل فيها مع محیطه الخارجي، أما الآليات الداخلية لا تظهر ضمن هذه المخططات.

#### 2 مخطط حالات الاستخدام الأولى

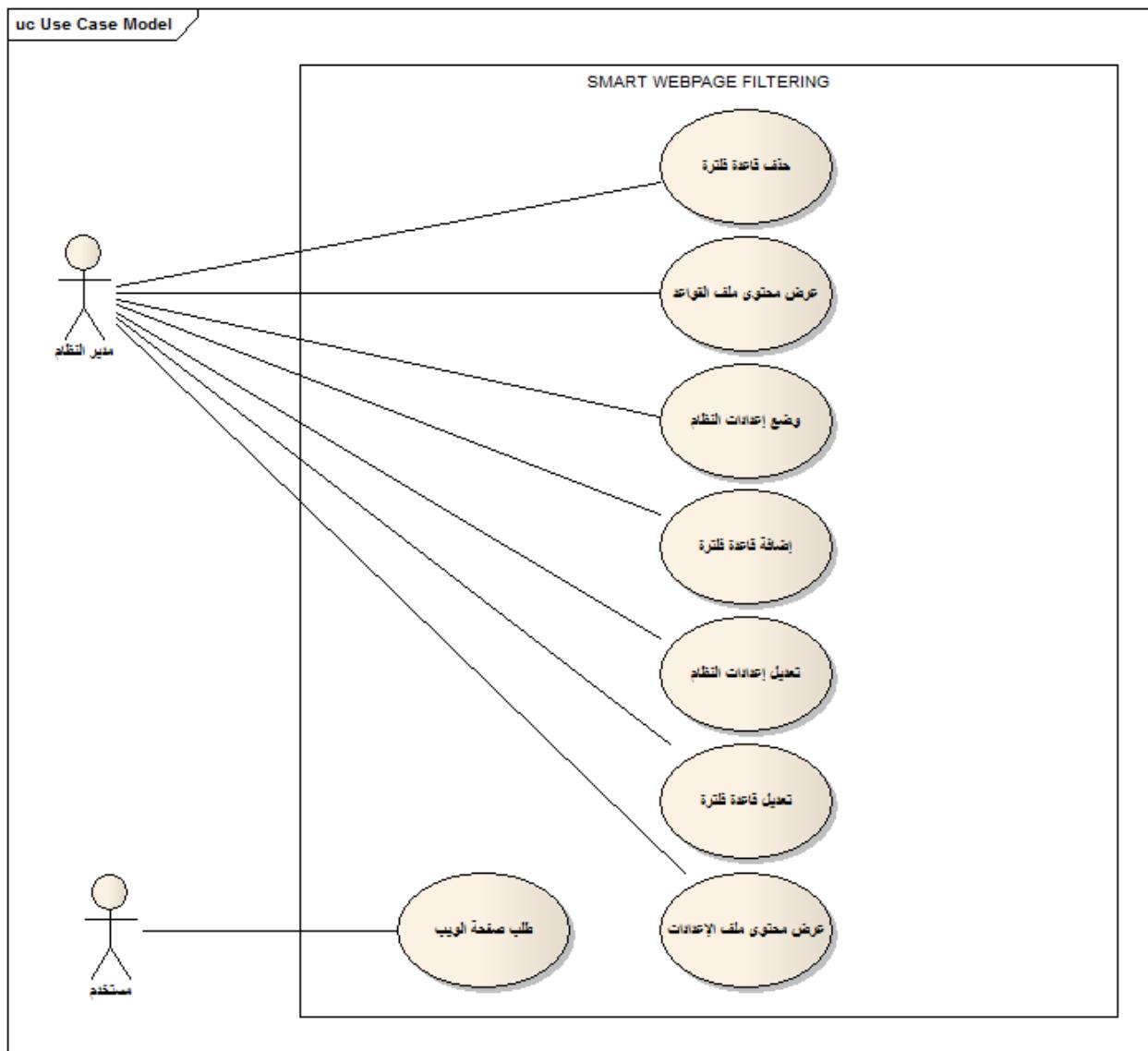
المخطط العام لحالات الاستخدام، يشمل حالات الاستخدام العامة دون الخوض في تفاصيلها الدقيقة، يعطي تصوراً مبدئياً عمّا مهمات النظام ومن هم الفاعلون المشاركون فيها.

## 2.1 مخطط حالات استخدام مخدم البروكسي Smaz Proxy



الشكل 44: مخطط حالات الاستخدام لمخدم البروكسي

## 2.2 مخطط حالات استخدام مصفّف الأخبار Smaz Filter



الشكل 4.5 مخطط حالات الاستخدام لقارئ الأخبار

### 3 حالات استخدام مخدم البروكسي Smaz Proxy

#### 3.1 وضع إعدادات النظام

حالات الاستخدام	تعيين إعدادات النظام
رقم حالة الاستخدام	1
الممثلين الأوليين	مدير النظام
الممثلين الثانويين	لا يوجد.
توصيف مختصر	تعيين إعدادات النظام والمعاملات الالزمة للعمل.
الشروط المسبقة	لا يوجد.
التدفق الأساسي للأحداث	<p>1- تبدأ حالة الاستخدام هذه عندما ي يريد مدير النظام وضع الإعدادات الخاصة بالنظام؛</p> <p>2- يقوم مدير النظام بتحديد الإعدادات من خلال وضع القيم الالزمة لكل إعداد ضابط للنظام:</p> <ul style="list-style-type: none"> <li>- عنوان خدمة الويب؛</li> <li>- البحث ضمن dbpedia؛</li> <li>- خوارزمية فك الغموض المتبعة؛</li> <li>- رد الصيائر إلى أصلها أم لا.</li> </ul>
الشروط اللاحقة	تم وضع إعدادات النظام من قبل مدير النظام.
طرق البديلة	لا يوجد.

جدول 5 وضع إعدادات النظام

### 3.2 تعديل إعدادات النظام

تعديل إعدادات النظام	حالة الاستخدام
2 رقم حالة الاستخدام	
مدير النظام	الممثلين الأوليين
لا يوجد.	الممثلين الثانويين
تعديل إعدادات النظام والمعاملات الالزمة للعمل.	توصيف مختصر
لا يوجد.	الشروط المسبقة
3- تبدأ حالة الاستخدام هذه عندما يريد مدير النظام تعديل الإعدادات الخاصة بالنظام؛	التدفق الأساسي للأحداث
4- يقوم مدير النظام بتعديل الإعدادات التي يريد من خلال وضع القيم الالزمة لكل إعداد ضابط للنظام:  - عنوان خدمة الويب؛ - إضافة الأوزان إلى الآباء؛ - البحث ضمن dbpedia؛ - خوارزمية فك الغموض المتبعة؛ - رد الضمائر إلى أصلها أم لا.	
تم تعديل إعدادات النظام من قبل مدير النظام.	الشروط اللاحقة
لا يوجد.	الطرق البديلة

جدول 6 تعديل إعدادات النظام

### 3.3 إضافة قاعدة فلترة

حالة الاستخدام	إضافة قاعدة فلترة.
رقم حالة الاستخدام	3
الممثلين الأوليين	مدير النظام
الممثلين الثانيين	لا يوجد.
توصيف مختصر	إضافة قاعدة إلى مجموعة القواعد التي سيتم على أساسها أخذ قرار الحجب أم لا.
الشروط المسبقة	لا يوجد.
التدفق الأساسي للأحداث	<p>1- تبدأ حالة الاستخدام هذه عندما يريد مدير النظام إضافة قاعدة فلترة إلى مجموعة القواعد المتواجدة؛</p> <p>2- يقوم مدير النظام بوضع كل مجال ونسبة ورودها ضمن الصفحة؛</p> <p>3- وضع الكلمات التي يريد ورودها ضمن الصفحة في حال تحقق الشرط الأول؛</p> <p>4- وضع مرادفات الكلمات التي يريد ورودها ضمن الصفحة؛</p> <p>5- وضع الحدث اللازم حدوثه عن تتحقق القاعدة وهو حجب الصفحة أم لا.</p>
الشروط اللاحقة	تم إضافة قاعدة فلترة إلى مجموعة القواعد المتواجدة مسبقا.
الطرق البديلة	لا يوجد.
لاستثناءات	لا يوجد.

جدول 7 إضافة قاعدة فلترة

### 3.4 تعديل قاعدة فلترة

حالة الاستخدام	تعديل قاعدة فلترة.
رقم حالة الاستخدام	4
الممثلين الأوليين	مدير النظام
الممثلين الثانيين	لا يوجد.
توصيف مختصر	تعديل قاعدة من قواعد الفلترة الموجودة مسبقا.
الشروط المسبقة	لا يوجد.
التدفق الأساسي للأحداث	<p>1- تبدأ حالة الاستخدام هذه عندما يريد مدير النظام تعديل قاعدة فلترة ضمن مجموعة القواعد المتواجدة؛</p> <p>2- في حال أراد تعديل المجالات يقوم مدير النظام بتعديل المجالات ونسبة ورودها ضمن الصفحة؛</p> <p>3- في حال أراد تعديل الكلمات يقوم المدير بتعديل الكلمات التي يريد ورودها ضمن الصفحة في حال تحقق الشرط الأول؛</p> <p>4- في حال أراد تعديل المرادفات يقوم بتعديل مرادفات الكلمات التي يريد ورودها ضمن الصفحة؛</p> <p>5- في حال أراد تعديل حدث القاعدة يقوم المدير بتعديل الحدث اللازم حدوثه عن تحقق القاعدة وهو حجب الصفحة أم لا.</p>
الشروط اللاحقة	تم تعديل قاعدة فلترة موجودة ضمن مجموعة القواعد المضافة مسبقا.

لا يوجد.	الطرق البديلة
لا يوجد.	لاستثناءات

جدول 8 تعديل قاعدة فلترة

### 3.5 طلب صفحة ويب

التدفق الأساسي	الشروط المسبقة	الممثلين الثانيين	الممثلين الأوليين	رقم حالة الاستخدام	طلب صفحة ويب	حالة الاستخدام
				5		
					مستخدم النظام	
					لا يوجد.	
					طلب صفحة ويب ومعالجتها.	توصيف مختصر
					لا يوجد.	الشروط المسبقة
1 - تبدأ حالة الاستخدام هذه عندما يريد مستخدم النظام طلب صفحة ويب ؛ 2 - يقوم المستخدم بطلب الصفحة التي يريد ؛ 3 - يمر هذا الطلب إلى squid proxy server ؛ 4 - بعدها يتم تمرير الطلب إلى icap server ليقوم بمعالجة الصفحة المرررة دلالياً ، 5 - يتم الحصول على نسبة كل مجال ضمن الصفحة ومقارنة النتيجة بالقواعد الموجودة والإقرار فيما اذا كانت الصفحة ستحجب أم لا ؛ 6 - في حال كان القرار هو حجب الصفحة يتم رجوع النتيجة بعكس مسار الطلب						

وعرض صفحة الحجب للمستخدم؛	
7- في حال كان القرار هو عدم حجب الصفحة يتم رجوع النتيجة بعكس مسار الطلب وعرض الصفحة للمستخدم؛	
تم طلب صفحة ويب ومعالجتها والإقرار فيما إذا كان سيتم عرض الصفحة أم لا.	الشروط اللاحقة
لا يوجد.	الطرق البديلة
لا يوجد.	لاستثناءات

**جدول 9 طلب صفحة ويب**

### 3.6 حذف قاعدة فلترة

حالة الاستخدام	حذف قاعدة فلترة.
رقم حالة الاستخدام	6
الممثلين الأوليين	مدير النظام
الممثلين الثانويين	لا يوجد.
توصيف مختصر	حذف قاعدة من قواعد الفلترة الموجودة مسبقا.
الشروط المسبقة	لا يوجد.
التدفق الأساسي للأحداث	1- تبدأ حالة الاستخدام هذه عندما يريد مدير النظام حذف قاعدة فلترة ضمن مجموعة القواعد المتواجدة؛ 2- يقوم المدير بتحديد القاعدة التي يريد حذفها؛ 3- يطلب المدير الأمر حذف القاعدة؛

الحذف.	4- يقوم النظام بحذف القاعدة وحفظ التعديلات الحاصلة في حال أكد المدير أمر
لا يوجد.	تم حذف قاعدة فلترة موجودة ضمن مجموعة القواعد المضافة مسبقا.
لا يوجد.	الشروط اللاحقة
لا يوجد.	الطرق البديلة
	لاستثناءات

جدول 10 حذف قاعدة فلترة

### 3.7 عرض محتوى ملف القواعد

حالة الاستخدام	عرض محتوى ملف القواعد.
رقم حالة الاستخدام	7
الممثلين الأوليين	مدير النظام
الممثلين الثانويين	لا يوجد.
توصيف مختصر	عرض محتوى ملف قواعد الحجب التي تم إضافتها مسبقا.
الشروط المسبقة	لا يوجد.
التدفق الأساسي للأحداث	<p>1- تبدأ حالة الاستخدام هذه عندما يريده مدير النظام عرض محتوى ملف القواعد والاطلاع على القواعد الموضوعة مسبقا؛</p> <p>2- يطلب المدير من النظام عرض ملف القواعد؛</p> <p>3- يتم عرض الملف ضمن نافذة تحوي القواعد بالترتيب الموجود مسبقا مع الحدث الذي يرافق كل قاعدة.</p>

تم عرض محتوى ملف القواعد الموضوعة مسبقاً من قبل مدير النظام.	الشروط اللاحقة
لا يوجد.	الطرق البديلة
لا يوجد.	لاستثناءات

جدول 11 عرض محتوى ملف القواعد

### 3.8 عرض محتوى ملف الإعدادات

عرض محتوى ملف الإعدادات	حالة الاستخدام
8	رقم حالة الاستخدام
مدير النظام	الممثلين الأوليين
لا يوجد.	الممثلين الثانيين
عرض محتوى ملف الإعدادات التي تم تحديدها مسبقاً.	توصيف مختصر
لا يوجد.	الشروط المسبقة
1- تبدأ حالة الاستخدام هذه عندما يريد مدير النظام عرض محتوى ملف الإعدادات والاطلاع على المحددات الموضوعة مسبقاً؛ 2- يطلب المدير من النظام عرض ملف الإعدادات؛ 3- يتم عرض الملف ضمن نافذة تحوي الإعدادات الموجودة ضمن الملف والتي تظهر بشكل إعداد مع قيمة هذا الإعداد.	التدفق الأساسي للأحداث
تم عرض محتوى ملف الإعدادات الموضوعة مسبقاً من قبل مدير النظام.	الشروط اللاحقة

لا يوجد.	الطرق البديلة
لا يوجد.	لاستثناءات

جدول 12 عرض محتوى ملف الإعدادات

#### 4 حالات استخدام إضافة المتصفح Smaz Filter

حالات استخدام إضافة المتصفح هي نفسها حالات استخدام مخدم البروكسي، الفرق الوحيد في هذا التطبيق أنه موضوع من أجل الأستخدامات الفردية، وأن طلب خدمة الويب التي تزودنا بالفهم الدلالي للصفحة تتم عن طريق الإضافة مباشرة ، بينما في المخدم كانت الخدمة يتم طلبها من خلال مخدم الويب.

#### 5 حالات استخدام مصنف الأخبار Smaz Reader

##### 5.1 عرض الأخبار التابعة لمجال معين

حالة الاستخدام	عرض الأخبار التابعة لمجال معين.
رقم حالة الاستخدام 1	
الممثلين الأوليين	مستخدم النظام
الممثلين الثانيين	لا يوجد.
توصيف مختصر	عرض الأخبار التابعة لمجال معين.
الشروط المسبقة	لا يوجد.
التدفق الأساسي للأحداث	1 - تبدأ حالة الاستخدام هذه عندما يريد مستخدم النظام عرض الأخبار التابعة لمجال معين؛

<p>2- يقوم النظام بعرض قائمة المجالات المتاحة؛</p> <p>3- يختار المدير المجال الذي يريد عرض أخباره؛</p> <p>4- يقوم النظام بالحصول على أخبار هذا المجال الموجودة ضمن قاعدة معطيات خاصة؛</p> <p>5- يقوم النظام بعرض الأخبار التابعة لهذا المجال.</p>	
تم عرض الأخبار التابعة لمجال معين من قبل مستخدم النظام.	الشروط اللاحقة
لا يوجد.	الطرق البديلة
لا يوجد.	لاستثناءات

**جدول 13 عرض الأخبار التابعة لمجال معين**

## 5.2 تحديد المجالات التي يرغب بعرض الأخبار التابعة لها

تحديد مجالات الأخبار المرغوب عرضها.	حالة الاستخدام
2	رقم حالة الاستخدام
مستخدم النظام	الممثلين الأوليين
لا يوجد.	الممثلين الثانويين
تحديد المجالات التي يرغب المستخدم بعرض الأخبار الخاصة بها.	توصيف مختصر
لا يوجد.	الشروط المسبقة
1- تبدأ حالة الاستخدام هذه عندما يريد مستخدم النظام تحديد المجالات التي يرغب بعرض الأخبار الخاصة بها؛	التدفق الأساسي للأحداث

<p>2- يقوم النظام بعرض قائمة تحوي كل المجالات المتاحة ؛</p> <p>3- يقوم مستخدم النظام بتحديد المجالات التي يريد ؛</p> <p>4- يتم حفظ التعديلات في حال أكد المستخدم ذلك.</p>	
تم تحديد المجالات التي يرغب المستخدم بعرض أخبارها.	الشروط اللاحقة
لا يوجد.	الطرق البديلة
لا يوجد.	لاستثناءات

جدول 14 تحديد المجالات التي يرغب بعرض الأخبار التابعة لها

### 5.3 وضع إعدادات التطبيق

حالة الاستخدام	وضع إعدادات التطبيق
رقم حالة الاستخدام	3
الممثلين الأوليين	مستخدم النظام
الممثلين الثانويين	لا يوجد.
توصيف مختصر	وضع إعدادات تطبيق الموبايل.
الشروط المسبقة	لا يوجد.
التدفق الأساسي للأحداث	<p>1- تبدأ حالة الاستخدام هذه عندما يريد مستخدم النظام وضع إعدادات تطبيق الموبايل الخاص بتصنيف الأخبار ؛</p> <p>2- يقوم المستخدم بتحديد درجة حساسية انتقاء الخبر لمجال معين ؛</p> <p>3- يقوم المستخدم بتحديد هل يريد إظهار عدد الأخبار ضمن كل مجال من</p>

المجالات ألم لا ؟	
-4 يطلب المستخدم من النظام حفظ الإعدادات.	
تم وضع إعدادات التطبيق من قبل مستخدم النظام.	الشروط اللاحقة
لا يوجد.	الطرق البديلة
لا يوجد.	لاستثناءات

جدول 15 وضع إعدادات التطبيق

#### 5.4 عرض المجالات المتواجدة

حالة الاستخدام	عرض المجالات المتواجدة.
رقم حالة الاستخدام	4
الممثلين الأوليين	مستخدم النظام
الممثلين الثانيين	لا يوجد.
توصيف مختصر	عرض المجالات المتواجدة والتي تحوي مجموعة من الأخبار.
الشروط المسбقة	لا يوجد.
التدفق الأساسي للأحداث	1- تبدأ حالة الاستخدام هذه عندما يريد مستخدم النظام معرفة المجالات الموجودة ضمن التطبيق؛
	2- يقوم النظام بعرض المجالات التي سبق وقد تم تحديدها من قبل المستخدم من بين كل المجالات المتوفرة.

تم عرض المجالات المرغوب فيها.	الشروط اللاحقة
لا يوجد.	الطرق البديلة
لا يوجد.	لاستثناءات

جدول 16 عرض المجالات المتواجدة

## 5.5 إضافة قناة أخبار

إضافة قناة أخبار.	حالة الاستخدام
5	رقم حالة الاستخدام
مستخدم النظام	الممثلين الأوليين
لا يوجد.	الممثلين الثانيين
إضافة قناة أخبار ليتم تحصيل الأخبار منها.	توصيف مختصر
لا يوجد.	الشروط المسبيقة
1- تبدأ حالة الاستخدام هذه عندما يريد مستخدم النظام إضافة قناة أخبار إلى قائمة القنوات المتواجدة؛ 2- يطلب المستخدم الأمر إضافة قناة؛ 3- يقوم المستخدم بإدخال عنوان القناة؛ 4- يقوم المستخدم بإدخال رابط صفحة الأخبار؛ 5- يقوم النظام عندها بتحصيل الأخبار الموجودة ضمن الصفحة ومعالجتها وتحديد المجالات التي ينتمي إليها كل خبر من الأخبار.	التدفق الأساسي للأحداث

تم إضافة قناة أخبار إلى قائمة القنوات الموجودة مسبقا.	الشروط اللاحقة
لا يوجد.	الطرق البديلة
لا يوجد.	لاستثناءات

جدول 17 إضافة قناة أخبار

## 5.6 عرض القنوات المتواجدة

عرض قنوات الأخبار الموجودة مسبقا.	حالة الاستخدام
6 رقم حالة الاستخدام	
مستخدم النظام	الممثلين الأوليين
لا يوجد.	الممثلين الثانويين
عرض قنوات الأخبار التي سبق وتم إضافتها ومعالجة الأخبار ضمنها.	توصيف مختصر
لا يوجد.	الشروط المسبقة
1- تبدأ حالة الاستخدام هذه عندما يريد مستخدم النظام عرض القنوات التي تم إضافتها مسبقا إلى التطبيق ؛ 2- يقوم النظام بالحصول على القنوات المتواجدة لديه ضمن قاعدة معطياته ؛ 3- يقوم النظام بعرض القنوات الموجودة للمستخدم.	التدفق الأساسي للأحداث
تم عرض القنوات المتواجدة مسبقا لمستخدم النظام.	الشروط اللاحقة
لا يوجد.	الطرق البديلة

لا يوجد.	لاستثناءات
----------	------------

جدول 18 عرض القنوات المتواجدة

### 5.7 حذف قناة

حالة الاستخدام	حذف قناة أخبار موجودة مسبقا.
رقم حالة الاستخدام	7
الممثلين الأوليين	مستخدم النظام
الممثلين الثانيين	لا يوجد.
توصيف مختصر	حذف قناة أخبار موجودة مسبقا.
الشروط المسبقة	لا يوجد.
التدفق الأساسي للأحداث	<p>1- تبدأ حالة الاستخدام هذه عندما يريد مستخدم النظام حذف قناة أخبار سبق وتم إضافتها؛</p> <p>2- يقوم النظام بعرض القنوات المتواجدة ضمن التطبيق؛</p> <p>3- يحدد المستخدم القناة التي يرغب بحذفها؛</p> <p>4- يقوم النظام بحذف القناة وبالتالي حذف كل الأخبار الموجودة ضمن هذه القناة.</p>
الشروط اللاحقة	تم حذف قناة أخبار من قبل مستخدم النظام.
الطرق البديلة	لا يوجد.
لاستثناءات	لا يوجد.

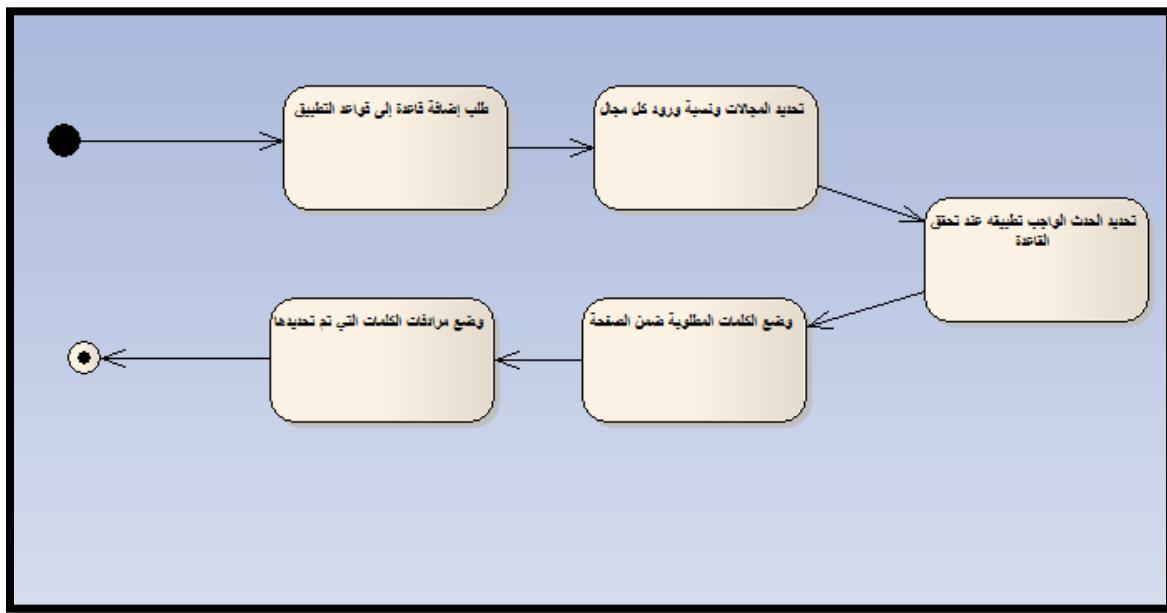
جدول 19 حذف قناة

## 5.8 وضع إعدادات النظام

تعيين إعدادات النظام	حالة الاستخدام
8	رقم حالة الاستخدام
مدير النظام	المثليين الأوليين
لا يوجد.	المثليين الثانويين
تعيين إعدادات النظام والمعاملات الالزمة للعمل.	توصيف مختصر
لا يوجد.	الشروط المسبقة
5- تبدأ حالة الاستخدام هذه عندما يزيد مدير النظام وضع الإعدادات الخاصة بالنظام؛	التدفق الأساسي لأحداث
6- يقوم مدير النظام بتحديد الإعدادات من خلال وضع القيم الالزمة لكل إعداد ضابط للنظام:  - عنوان خدمة الويب؛ - رفع الأوزان إلى الآباء؛ - البحث ضمن dbpedia؛ - خوارزمية فك الغموض المتبعة؛ - رد الضمائر إلى أصلها أم لا.	
تم وضع إعدادات النظام من قبل مدير النظام.	الشروط اللاحقة
لا يوجد.	الطرق البديلة

**جدول 20 وضع إعدادات النظام**

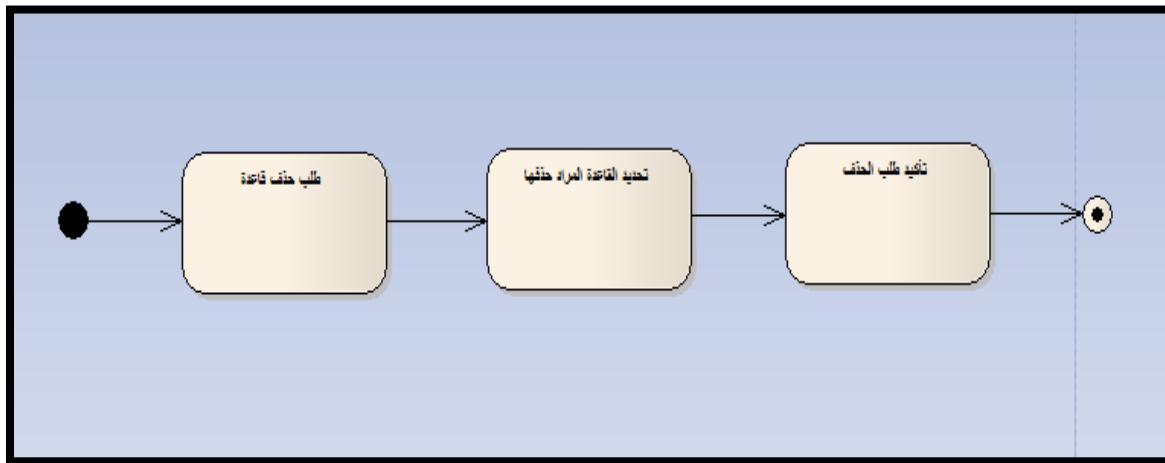
### إضافة قاعدة إلى قواعد الفلترة



الشكل 46 إضافة قاعدة إلى قواعد الفلترة

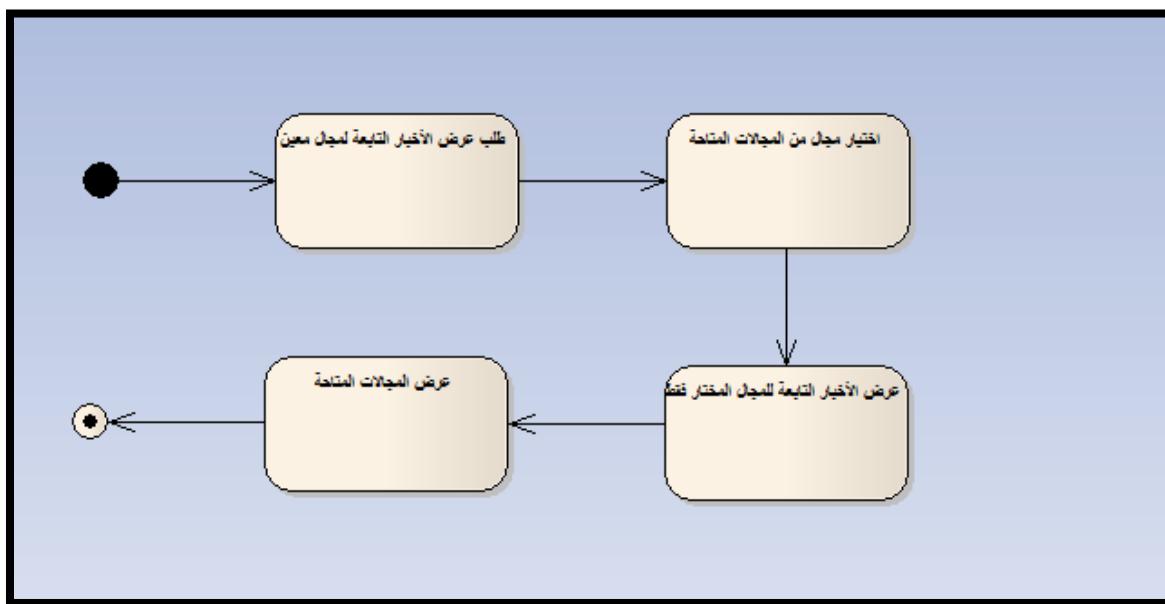


## حذف قاعدة



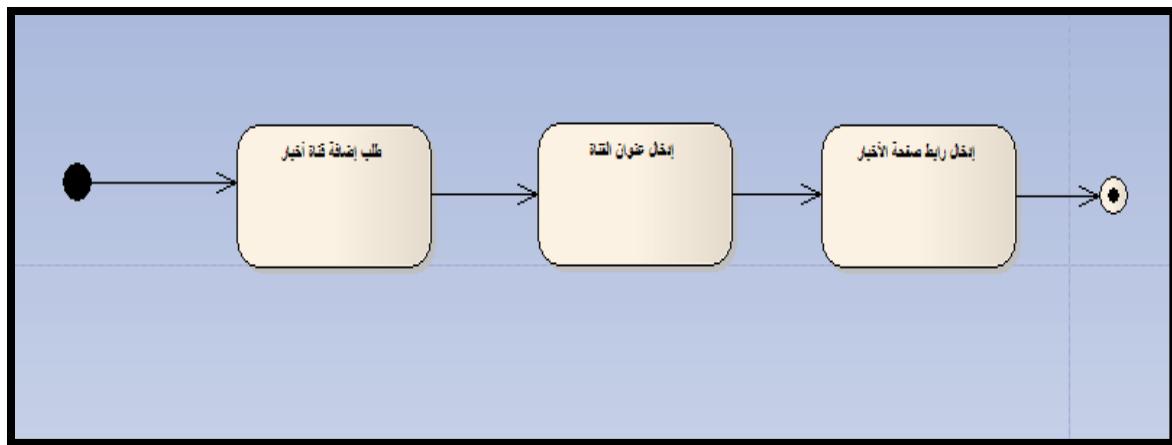
الشكل 47 حذف قاعدة

## عرض الأخبار التابعة لمجال معين



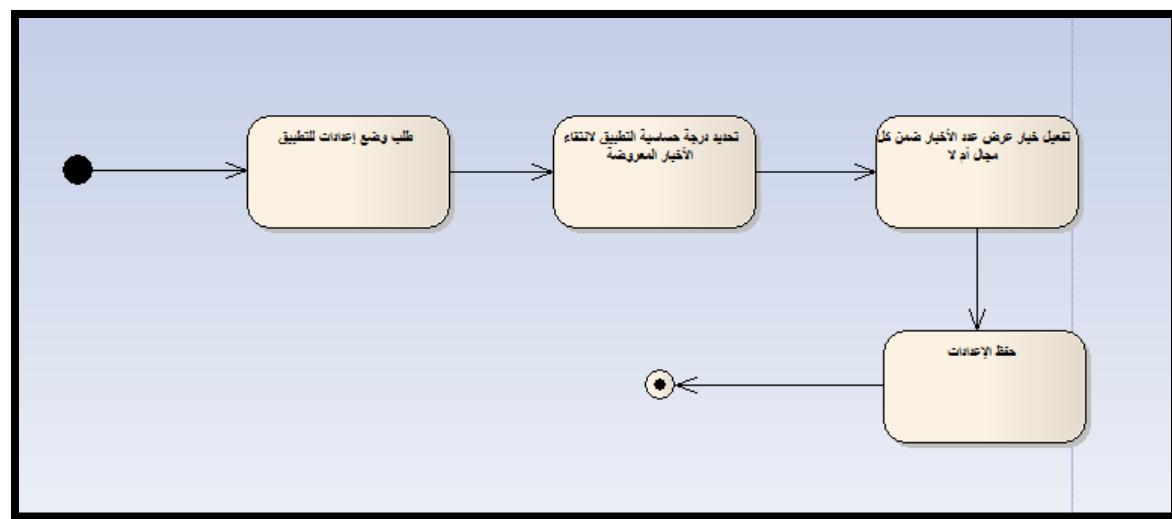
الشكل 48 عرض الأخبار التابعة لمجال معين

## إضافة قناة أخبار



الشكل 49 إضافة قناة أخبار

## وضع إعدادات التطبيق



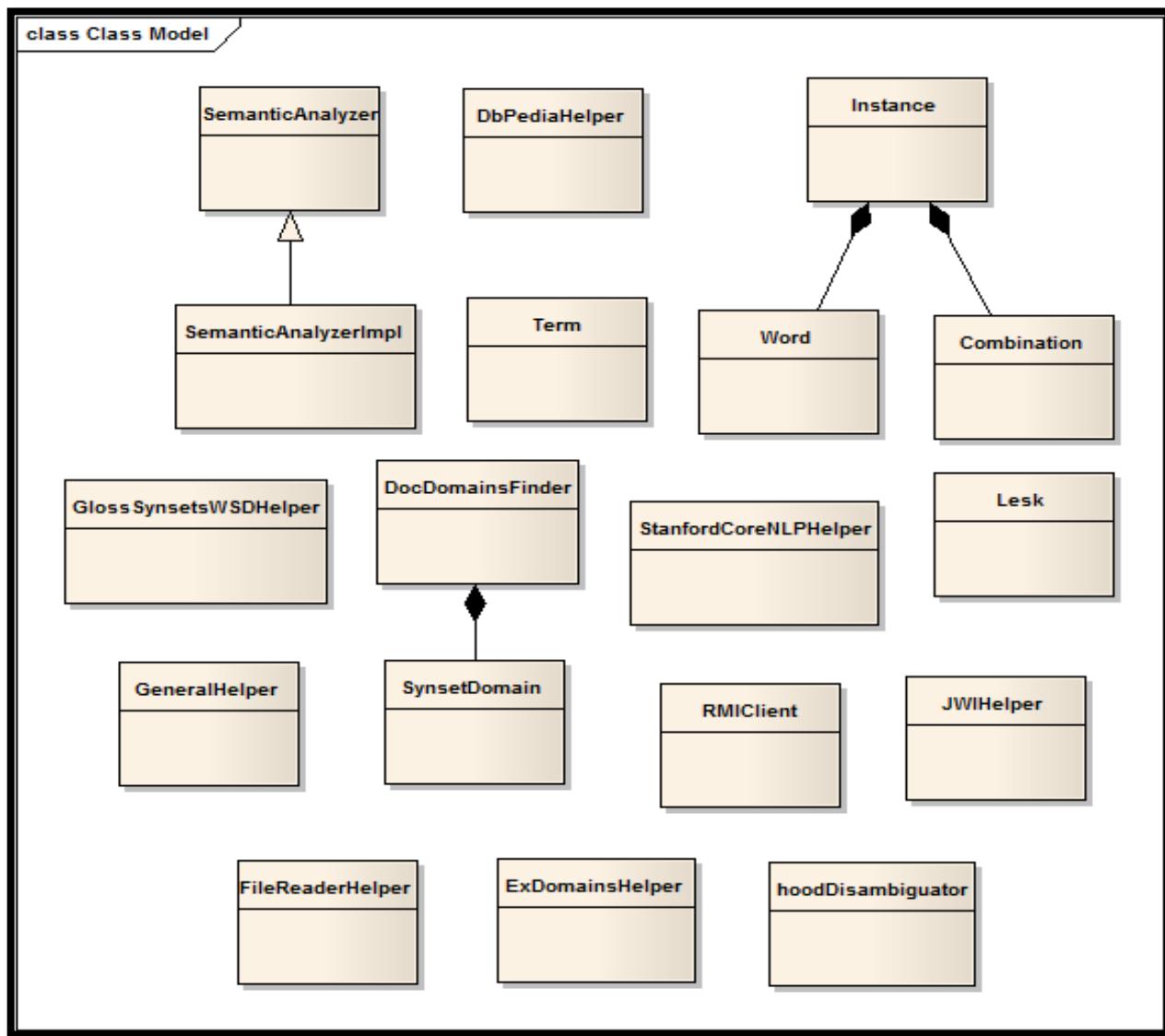
الشكل 50 وضع إعدادات للتطبيق

الفصل التاسع

## مذلّل الصّفوف المفاهيمي

---

## 1 مخطط الصنوف المفاهيمي لخدم البروكسي Smaz Proxy



الشكل 1.5 مخطط الصنوف المفاهيمي لخدم البروكسي

## 2 توصيف الصنوف المفاهيمية لخدم البروکسی

التصنيف	الصنف
<ul style="list-style-type: none"> <li>- واجهة خدمة الويب التي تقدم خدماتين أساسيتين هما تحميل المكتبات ومعالجة طلب فهم محتوى.</li> </ul>	<b>SemanticAnalyzer</b>
<ul style="list-style-type: none"> <li>- تحقيق الواجهة السابقة يعمل هذا الصنف على إعادة تعريف الخدمات الموجودة في الصنف السابق وتحقيقها؛</li> <li>- يشكل هذا الصنف الخدمة التي يتم الاستعانة بها من أجل فهم المحتوى.</li> </ul>	<b>SemanticAnalyzerImpl</b>
<ul style="list-style-type: none"> <li>- يعمل هذا الصنف كزبون للخدمة السابقة والذي يقوم بطلب معالجة محتوى وفهمه من الخدمة.</li> </ul>	<b>RMIClient</b>
<ul style="list-style-type: none"> <li>- هذا الصنف يقوم بإيجاد جميع الترکيبات الممكنة من معانی الكلمات الموجودة في النافذة ويقوم بالمرور على كل تركيبة وحساب رصيدها، ومن ثم اختيار التركيبة ذات الرصيد الأعلى لتكون كلمات هذه التركيبة هي المعانی الصحيحة للكلمات الموجودة في النافذة</li> </ul>	<b>Instance</b>

<p>- هذه الصفة يحوي مجموعة من الكلمات بحجم النافذة المحددة وبمعنى محدود ليتم فحص توافقها مع بعضها البعض لإيجاد الرصيد (score) النهائي لهذه التركيبة والذي يعبر عن مدى ارتباط كلماتها ببعضها البعض.</p>	<p><b>Combination</b></p>
<p>- هذا الصف يعبر عن الكلمة حيث يتم تخزين فيه موقع هذه الكلمة في الجملة ومعاني هذه الكلمة جميعها بالإضافة إلى أفضل معنى لها (سينتاج معنا من الخوارزمية).</p>	<p><b>Word</b></p>
<p>- إزالة غموض الكلمات من خلال خوارزمية lesk</p>	<p><b>Lesk</b></p>
<p>- يعمل هذا الصف على فك غموض جملة من خلال الاستعانة بخوارزمية hood في إزالة غموض الكلمات.</p>	<p><b>hoodDisambiguator</b></p>
<p>- يشكل هذا الصف مفهوم كلمة من الكلمة بحد ذاتها ومكانها في الجملة والجذر الخاص بها.</p>	<p><b>Term</b></p>
<p>- يقدم هذا الصف عمليات مساعدة من أجل التعامل مع مكتبة StanfordCoreNLP</p>	<p><b>StanfordCoreNLPHelper</b></p>

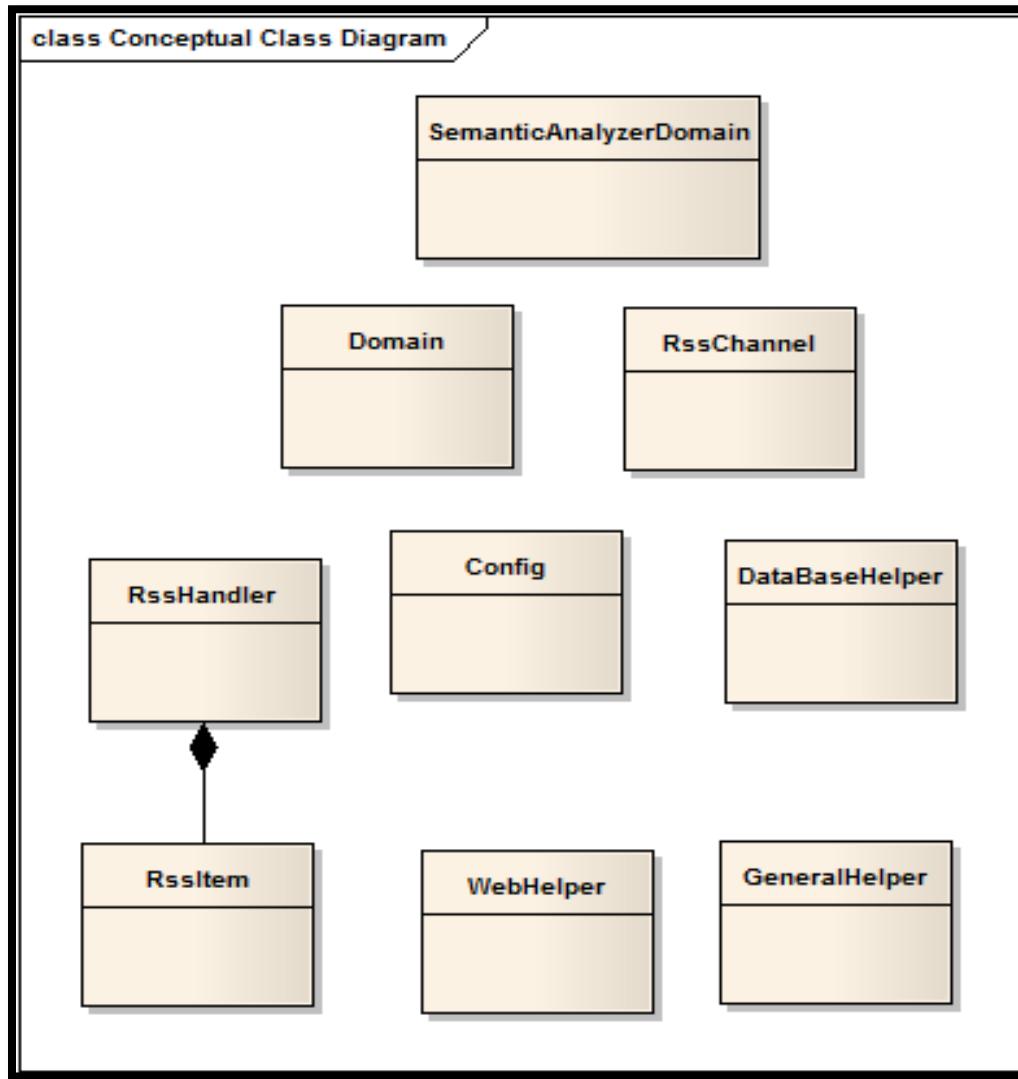
<p>مثل الحصول على أصل الكلمة.</p>	
<p>- يقدم هذا الصنف عمليات مساعدة من أجل التعامل مع مكتبة wordnet مثل الحصول على المجموعة الأشهر لكلمة، والحصول على المجموعات التي تنتهي لها كلمة ما.</p>	<b>JWIHelper</b>
<p>- يقدم هذا الصنف عمليات مساعدة من أجل التعامل مع أنطولوجية dbpedia مثل الاستعلام ضمنها والحصول على معنى كلمة ما.</p>	<b>DbpediaHelper</b>
<p>- يقدم هذا الصنف عمليات مساعدة من أجل التعامل مع wordnet domains مثل الحصول على آباء مجال ما والحصول على المجالات الموجودة وتحميلها مع الأوزان إلى الذاكرة.</p>	<b>ExDomainsHelper</b>
<p>- يقدم هذا الصنف عمليات من أجل قراءة محتوى word, pdf, png, html ....</p>	<b>FileReaderHelper</b>
<p>- يقدم هذا الصنف عمليات مساعدة من أجل خوارزمية إزالة الغموض المتعلقة ب gloss</p>	<b>GlossSynsetsWSDHelper</b>

<b>.synset</b>	
<p>- يقدم هذا الصنف عمليات مساعدة من أجل عمليات معالجة اللغات الطبيعية من تحميل الكلمات التي لا معنى لها وتقسيم الجمل والكلمات.</p>	<b>GeneralHelper</b>
<p>- يقدم هذا الصنف عمليات مساعدة من أجل التعامل مع domains مثل إسناد الاسم والوزن لكل مجال ضمن مجموعة ما.</p>	<b>SynsetDomain</b>
<p>- يقدم هذا الصنف عمليات مساعدة من أجل التعامل مع domains مثل رفع الأوزان إلى الآباء وعمل تنظيم لقيم الأوزان ضمن مجال محدد.</p>	<b>DocDomainsFinder</b>

جدول 21 توصيف الصنوف المفاهيمية لمخدم البروكسي

### 3 مخطط الصنف المفاهيمي للقارئ Smaz Reader

- هنا تتكرر نفس الصنف التي لها علاقة بالفهم الدلالي للمحتوى لذلك لم يتم إعادة رسمها؛
- تم وضع الصنف التي لها علاقة بتطبيق الموبايل الخاص بتصنيف الأخبار.



الشكل 52 مخطط الصنف المفاهيمي لقارئ الأخبار

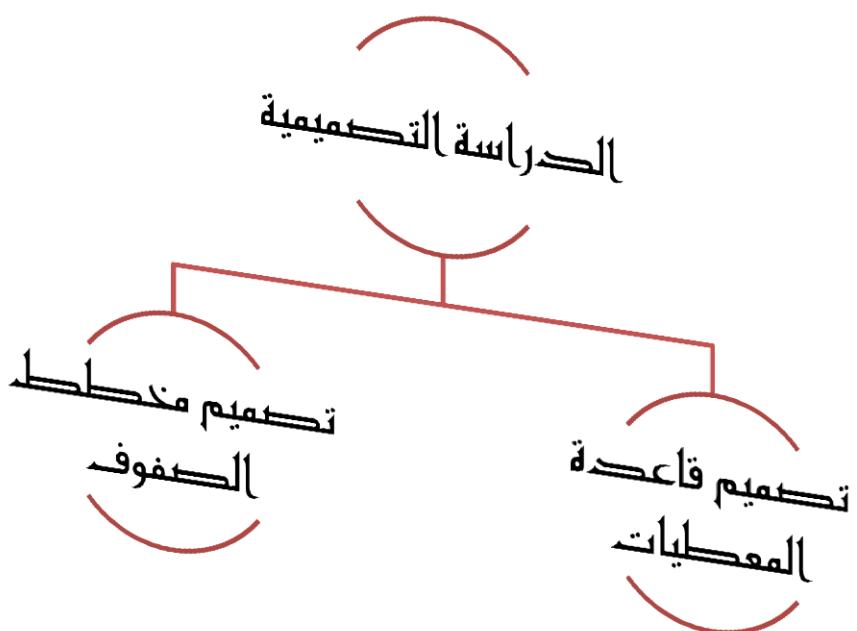
#### 4 توصيف الصفوف المفاهيمية للقارئ

التصنيف	الصف
<ul style="list-style-type: none"> <li>- يؤمن هذا الصف مجال معين موجود ضمن محتوى ما ونسبة وجوده ضمن المحتوى؛</li> <li>- عند معالجة محتوى ما يتم الحصول على قائمة من هذا الصف.</li> </ul>	<b>SemanticAnalyzerDomain</b>
<ul style="list-style-type: none"> <li>- يمثل هذا الصف مجال معين.</li> </ul>	<b>Domain</b>
<ul style="list-style-type: none"> <li>- يمثل هذا الصف خبر من الأخبار الموجودة ضمن قناة الأخبار.</li> </ul>	<b>RssItem</b>
<ul style="list-style-type: none"> <li>- يمثل هذا الصف قناة الأخبار.</li> </ul>	<b>RssChannel</b>
<ul style="list-style-type: none"> <li>- يؤمن هذا الصف مجموعة من العمليات المساعدة للتعامل مع قواعد المعطيات من إضافة وحذف وتعديل للمجالات والقنوات والأخبار.</li> </ul>	<b>DataBaseHelper</b>
<ul style="list-style-type: none"> <li>- يؤمن هذا الصف مجموعة من العمليات المساعدة مثل تحميل المجالات وإضافتها إلى القائمة.</li> </ul>	<b>GeneralHelper</b>
<ul style="list-style-type: none"> <li>- يؤمن هذا الصف مجموعة من العمليات المساعدة مثل الاتصال بالإنترنت وتحميل الأخبار الموجودة ضمن قناة ما.</li> </ul>	<b>WebHelper</b>

جدول 22 توصيف الصفوف المفاهيمية للقارئ

## الباب الرابع

### الدراسة التصميمية



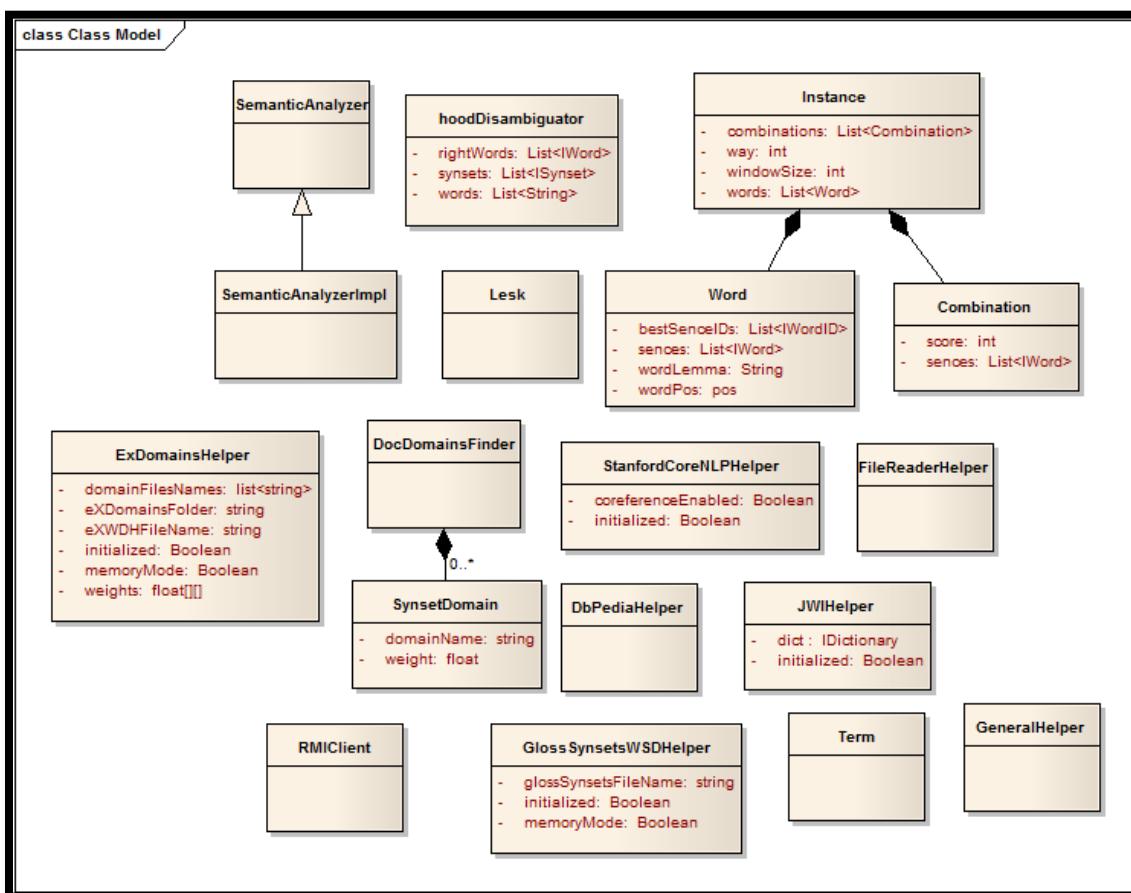
الشكل 53 النقاط الأساسية ضمن مرحلة التصميم



# الفصل العاشر

## مخطط صنف مخدود البروکسی

1 مخطط صنف مخدود البروکسی Smaz Proxy



الشكل 54 مخطط صنف مخدود البروکسی

## 2 شرح بعض صفات مخدم البروكسي Smaz Proxy

### hoodDisambiguator 2.1

hoodDisambiguator
- rightWords: List<IWord> - synsets: List<ISynset> - words: List<String>

- يستخدم هذا الصنف من أجل إزالة غموض جملة عن طريق خوارزمية إزالة الغموض، hood يحتوي على الخصائص التالية:

- سلسلة الكلمات المقابلة بعد إزالة غموض الكلمات؛
  - سلسلة المجموعات التي سنحصل عليها عند انتهاء الخوارزمية من عملها، حيث ترد الخوارزمية المجموعة التي تنتمي إليها كل كلمة من الكلمات التي ترغب بإزالة غموضها؛
  - سلسلة من الكلمات التي ترغب بإزالة غموضها.
- يحتوي العمليات التالية:
  - الحصول على الكلمات من الجملة المرررة؛
  - إزالة غموض الكلمات؛
  - إيجاد الجذر HoodRoot؛
  - إيجاد كل الآباء التابعين لمجموعة ما؛
  - الحصول على أكثر مجموعة مشهورة لكلمة ما.

## Word 2.2

Word
- bestSenoIDs: List<IWordID> - senos: List<IWord> - wordLemma: String - wordPos: pos

- هذا الصف يعبر عن الكلمة حيث يتم تخزين فيه موقع هذه الكلمة في الجملة ومعاني هذه الكلمة جميعها بالإضافة إلى أفضل معنى لها (سينتاج معنا من الخوارزمية).

## Combination 2.3

Combination
- score: int - senos: List<IWord>

- هذه الصف يحوي مجموعة من الكلمات بحجم النافذة المحددة وبمعاني محددة، ليتم فحص تواف قها مع بعضها البعض لإيجاد الرصيد (score) النهائي لهذه التركيبة والذي يعبر عن مدى ارتباط كلماتها ببعضها البعض.

- تتم عملية حساب الرصيد (score) من خلال التابع calculate\_Score الذي يقوم بتشكيل جميع الأزواج الممكنة من الكلمات الموجودة في التركيبة وايجاد التشابهات بين معاني كلمات الزوج الواحد وحساب عدد التشابهات مع الأخذ بعين الاعتبار العلاقات المرتبطة بكل كلمة والطريقة المحددة لحساب الأزواج (Heterogeneous, مع الأخذ بعين الاعتبار العلاقات المرتبطة بكل كلمة والطريقة المحددة لحساب الأزواج (Homogeneous).

## Instance 2.4

Instance
- combinations: List<Combination> - way: int - windowSize: int - words: List<Word>

- هذا الصف يقوم بإيجاد جميع التركيبات الممكنة من معاني الكلمات الموجودة في النافذة ويقوم بالمرور على كل تركيبة وحساب رصيدها، ومن ثم اختيار التركيبة ذات الرصيد الأعلى لتكون كلمات هذه التركيبة هي المعاني الصحيحة للكلمات الموجودة في النافذة.

- يحوي العمليات التالية:

الحصول على كل التراكيب المكونة ؛

الحصول على أفضل تركيبة.

### DbpediaHelper 2.5

- يقدم هذا الصنف عمليات مساعدة من أجل التعامل مع أنطولوجية dbpedia مثل الاستعلام ضمنها والحصول على

معنى كلمة ما.

- يحوي العمليات التالية :

طلب خدمة من spotlight والحصول على نتيجة منها؛

الحصول على روابط كلمات النص المدخل ضمن dbpedia،

استعلام sparql ضمن .dbpedia

### ExDomainsHelper 2.6

- يقدم هذا الصنف عمليات مساعدة من أجل التعامل مع .wordnet domains

ExDomainsHelper
- domainFileNames: list<string>
- eXDomainsFolder: string
- eXWDHFileName: string
- initialized: Boolean
- memoryMode: Boolean
- weights: float[][]

- يحوي العمليات التالية :

تحميل الأوزان إلى الذاكرة؛

الحصول على آباء مجال ما؛

الحصول على المجالات المتواجدة.

## FileReaderHelper 2.7

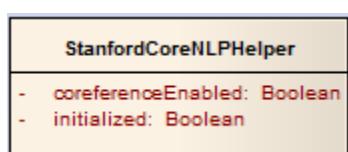
- يقدم هذا الصنف عمليات من أجل قراءة محتوى بصيغ مختلفة سواء كان .... word, pdf, html
- يقدم هذا الصنف العمليات التالية:
  - قراءة ملف pdf;
  - قراءة ملف DOCX;
  - قراءة ملف html;
  - الحصول على نص من ملف html;
  - الحصول على محتوى تاغات أخرى موجودة ضمن صفحة html.

## GeneralHelper 2.8

- يقدم هذا الصنف عمليات مساعدة من أجل عمليات معالجة اللغات الطبيعية من تحميل الكلمات التي لا معنى لها وتقسيم الجمل والكلمات.
- بعض العمليات التي يقدمها:
  - فصل الكلمات المتراكبة;
  - إزالة المحارف الغريبة;
  - تحميل الكلمات الغير حاملة للمعنى؛
  - الحصول على الكلمات الموجودة ضمن الجمل.

## StanfordCoreNLPHelper 2.9

- يقدم هذا الصنف عمليات مساعدة من أجل التعامل مع مكتبة StanfordCoreNLP مثل الحصول على أصل الكلمة.



- بعض العمليات التي يقدمها:

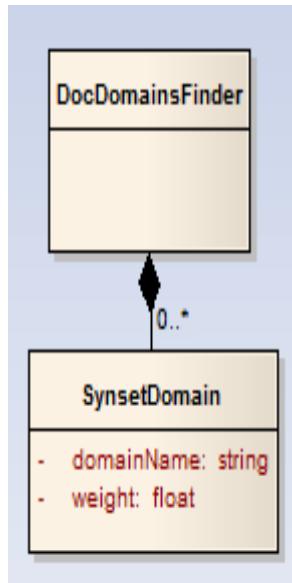
الحصول على أصل الكلمة؛

الحصول على الكلمات الموجودة ضمن موقع مقبولة ومحددة مسبقاً.

## DocDomainsFinder 2.10

- يقدم هذا الصنف عمليات مساعدة من أجل التعامل مع domains مثل رفع الأوزان إلى الآباء وعمل تنظيم لقيم الأوزان ضمن مجال محدد.

- بعض العمليات التي يقدمها:



إضافة وزن الكلمة إلى مجموعة المجال الكلي؛

إضافة الوزن إلى الآباء؛

عمل تنظيم للأوزان ضمن مجال محدد؛

الحصول على المجالات والأوزان الخاصة بها من أجل محتوى ما.

## SynsetDomain 2.11

- يقدم هذا الصنف عمليات مساعدة من أجل التعامل مع domains مثل إسناد الاسم والوزن لكل مجال ضمن مجموعة ما.

## JWIHelper 2.12

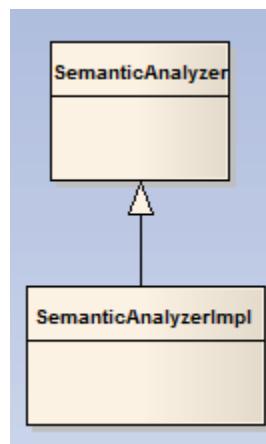
- يقدم هذا الصنف عمليات مساعدة من أجل التعامل مع مكتبة wordnet،

- بعض العمليات التي يقدمها:

- مثل الحصول على المجموعة الأشهر لكلمة،

- الحصول على المجموعات التي تنتهي لها كلمة ما

## SemanticAnalyzer 2.13



- واجهة خدمة الويب؛

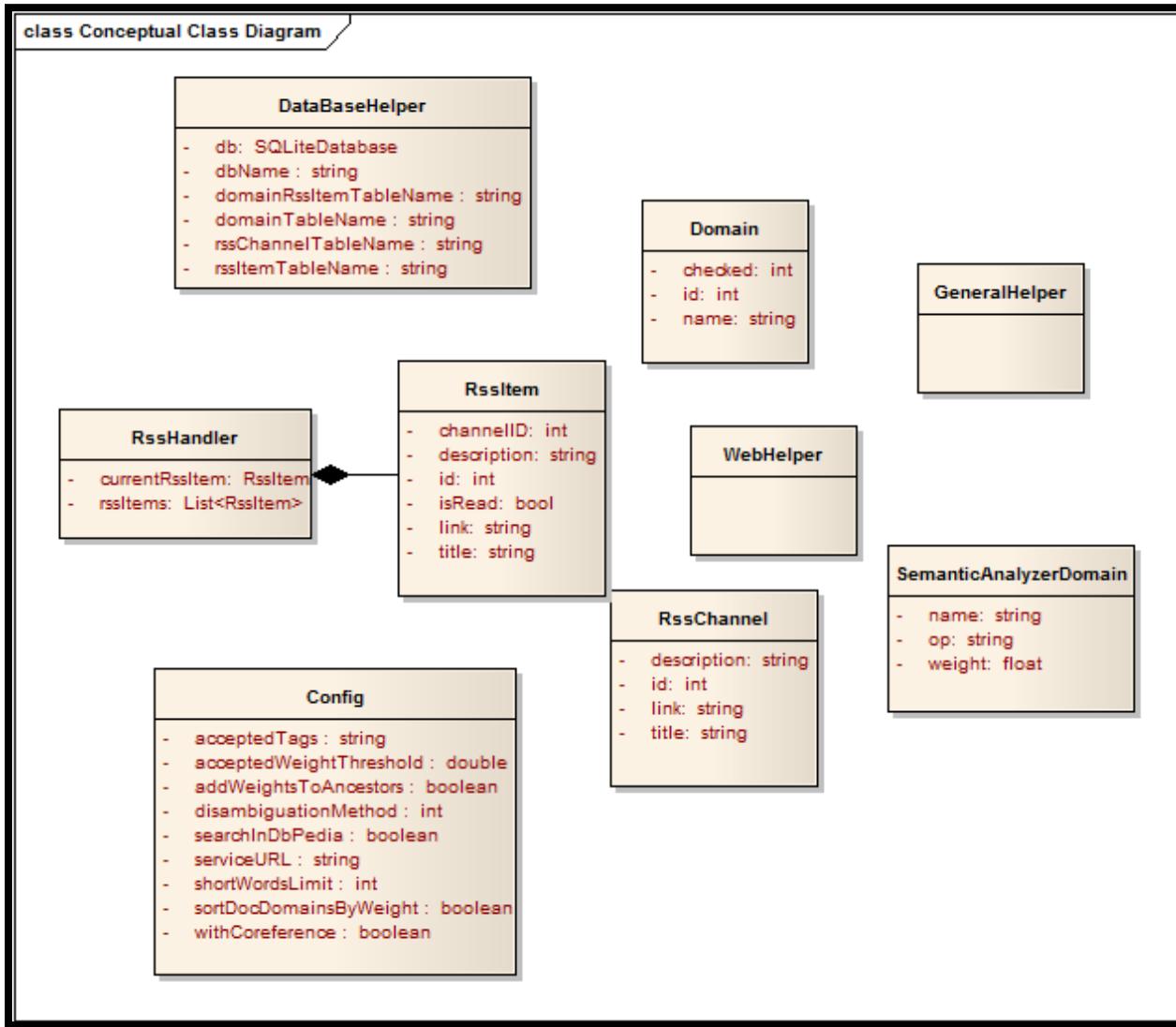
- تقدم العمليتين الأساسية التاليتين:

- تحميل المكتبات إلى الذاكرة؛

- معالجة طلب فهم محتوى.

## SemanticAnalyzerImpl 2.14

- تحقيق الواجهة السابقة يعمل هذا الصنف على إعادة تعريف الخدمات الموجودة في الصنف السابق وتحقيقها؛
- يشكل هذا الصنف الخدمة التي يتم الاستعانة بها من أجل فهم المحتوى.



الشكل 5 مخطط صنوف قارئ الأخبار

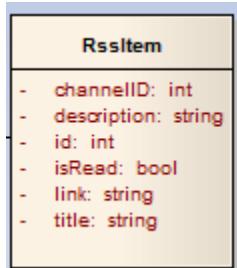
## 4 شرح بعض صفات القارئ Smaz Reader

### RssChannel 4.1



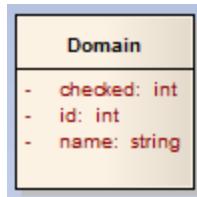
- يمثل هذا الصنف قناة أخبار؛
- يحوي هذا الصنف الخصائص التالية:
  - توصيف قناة الأخبار؛
  - المعرف المميز لقناة الأخبار؛
  - رابط وجود صفحة الأخبار؛
  - عنوان قناة الأخبار.

### RssItem 4.2



- يمثل هذا الصنف خبر من الأخبار الموجودة ضمن قناة الأخبار؛
- يحوي هذا الصنف الخصائص التالية:
  - توصيف الخبر؛
  - المعرف المميز للخبر؛
  - رابط وجود الخبر؛
  - قناة الأخبار التي يتبع لها الخبر؛
  - هل تم قراءة الخبر من قبل أم لا؛
  - عنوان الخبر.

### Domain 4.3



- يمثل هذا الصف مجال معين؛

- يحوي الخصائص التالية:

هل تم تحديد المجال من قبل أم لا؟

المعرف المميز للمجال؛

اسم المجال.

### DataBaseHelper 4.4

- يؤمن هذا الصف مجموعة من العمليات المساعدة للتعامل مع قواعد المعطيات؛

- يحوي العمليات التالية:

إضافة مجال؛

إضافة كل المجالات؛

الحصول على كل المجالات؛

الحصول فقط على المجالات التي تم تحديدها؛

تعديل مجال ما؛

إضافة قناة أخبار؛

تعديل قناة أخبار؛

حذف قناة أخبار؛

الحصول على كل قنوات الأخبار المتواجدة؛

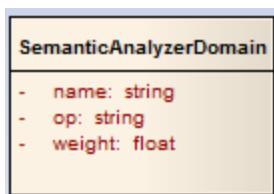
إضافة خبر؛

تحديد المجال الخاص بخبر ما؛

الحصول على كل الأخبار؛

الحصول على الأخبار التابعة لمجال معين.

## SemanticAnalyzerDomain 4.5



- يؤمن هذا الصنف مجال معين موجود ضمن محتوى ما ونسبة وجوده ضمن

المحتوى؛

- عند معالجة محتوى ما يتم الحصول على قائمة من هذا الصنف؛

- يحوي الخصائص التالية:

اسم المجال؛

العملية المطبقة على الوزن الخاص بالمجال؛

قيمة الوزن.

## WebHelper 4.6

- يؤمن هذا الصنف مجموعة من العمليات المساعدة مثل الاتصال بالانترنت وتحميل الأخبار الموجودة ضمن قناة ما.

- بعض العمليات التي يقدمها هذا الصنف:

تحميل كل الأخبار الموجودة ضمن كل القنوات المطلوبة؛

معالجة كل خبر من الأخبار ومعرفة المجالات التي ينتمي إليها.

## Config 4.7

Config
- acceptedTags : string - acceptedWeightThreshold : double - addWeightsToAncestors : boolean - disambiguationMethod : int - searchInDbPedia : boolean - serviceURL : string - shortWordsLimit : int - sortDocDomainsByWeight : boolean - withCoreference : boolean

- يمثل هذا الصنف مجموعة المتحولات التي يتم ضبط قيمها عند بداية عمل النظام؛

- يحوي الخصائص التالية:

▪ serviceURL عنوان خدمة الويب التي سيتم الاتصال بها والاستفادة من

خدماتها؛

▪ sortDocDomainsByWeight متاحول منطقي يأخذ قيمة true عندما نريد ترتيب

domains الصفحة تنازلياً حسب نسبة وردهم ضمن الصفحة؛

▪ addWeightsToAncestors متاحول منطقي يأخذ قيمة true عندما نريد رفع وزن الكلمة التابعة

ل domain ما إلى آباء هذا ال domain فعادة عند ورود كلمة تنتهي إلى domain ما بنسبة 0.5

فتعطي حرية رفع هذا الوزن إلى آباء هذا ال domain أم لا من خلال هذا المتاحول المنطقي؛

▪ shortWordsLimit متاحول صحيح يدل على الطول الأصغرى للكلمة التي يجب أخذها في الحساب

إذا كانت قيمته مثلاً 3 فعندها كل كلمة طولها أصغر من 3 لن يتم أخذها بعين الاعتبار عند المعالجة؛

▪ searchInDbPedia متاحول منطقي يأخذ قيمة true عندما نريد البحث عن الكلمة غير موجودة ضمن

ال wordnet ضمن ال dbpedia وفي حال لم تكن الكلمة موجودة ضمن ال wordnet وكان هذا

المتاحول المنطقي true عندها سيتم البحث عن هذه الكلمة ضمن ال dbpedia؛

▪ dbpediaResultsLimit متاحول صحيح يدل على عدد القيم التي سيتم أخذها بعين الاعتبار من أجل

كل علاقة سيتم الاستفادة منها من ال dbpedia، بمعنى أنه عندما لا نجد الكلمة ضمن ال wordnet

سيتم البحث عنها ضمن ال dbpedia، صفحة ال dbpedia تحوي مجموعة من العلاقات التي من

الممكن الاستفادة منها في محاولة لفهم معنى الكلمة، من هذه العلاقات هي علاقة subject، فمن أجل هذه

العلاقة ومن أجل الكلمة المطلوبة هناك مجموعة من القيم التي تشرح معنى هذه الكلمة وبالتالي لابد من تحديد العدد الأعظمي للقيم التي سيتمأخذها بعين الاعتبار؛

disambiguationMethod متحول صحيح يدل على الخوارزمية المتّبعة لإزالة الغموض فكما ذكرنا ▪

سابقاً كنا قد حققنا مجموعة من الخوارزميات لإزالة غموض كلمة كل منها مختلفة في تعقيدها وسرعتها ودققتها عن الأخرى

- الرقم 1 يعني أن الخوارزمية المتّبعة هي خوارزمية المعنى الأكثر شهرة؛
- WordNet Semantically Tagged يعني أن الخوارزمية المتّبعة هي خوارزمية glosses؛
- الرقم 3 يعني أن الخوارزمية المتّبعة هي خوارزمية Hood؛
- الرقم 4 يعني أن الخوارزمية المتّبعة هي خوارزمية Lesk.

acceptedTags متحول للدلالة على نوع الكلمات التي نرغب بمعالجتها ضمن النص سواء كانت فعل، اسم، ▪

صفة أو ظرف؛

withCoreference متحول منطقي يأخذ قيمة true عندما نريد إعادة الفحص إلى أصلها أثناء معالجة ▪

النص المطلوب؛

dbpediaPredicates متحول لتحديد العلاقات التي نريد معالجتها والبحث فيها ضمن أنطولوجية ▪

dbpedia searchInDbpedia عندما يكون خيار dbpedia مفعّل؛

leskWindowSize متحول لتحديد حجم النافذة الأصغر الذي نريد أن نعالجه ضمن خوارزمية إزالة الغموض ▪

lesk الرابعة؛

leskComparsionWay lesk متحول لتحديد طريقة المقارنة في خوارزمية ▪



# الفصل الحادي عشر

## تصميم قاعدة البيانات

1 جداول توصيف الكيانات

RssChannel		
النط	الشرح	الواصفة
int	رقم قناة الأخبار المميز	Id
varchar(50)	عنوان القناة	Title
varchar(50)	رابط القناة	Link
varchar(50)	توصيف القناة	Description

جدول 23 توصيف الكيان rssChannel

Domain		
النط	الشرح	الواصفة
int	رقم المجال المميز	Id
varchar(50)	اسم المجال	Name
int	تحديد المجال	Checked

جدول 24 توصيف الكيان Domain

RssItem		
النط	الشرح	الواصفة
int	رقم الخبر المميز	Id
varchar(50)	عنوان الخبر	Title
varchar(50)	رابط الخبر	Link
varchar(50)	توصيف الخبر	Description
Int	قراءة الخبر	isRead
Int	القناة التي يتبع لها الخبر	channelID

جدول 25 توصيف الكيان RssItem

DomainRssItem		
النوع	الشرح	الواصفة
int	الرقم المميز	Id
Int	رقم المجال الذي يتبع له الخبر	domainID
Int	رقم الخبر	RssItemID
float	وزن الخبر بالنسبة للمجال	Weight

جدول 26 توصيف الكيان DomainRssItem

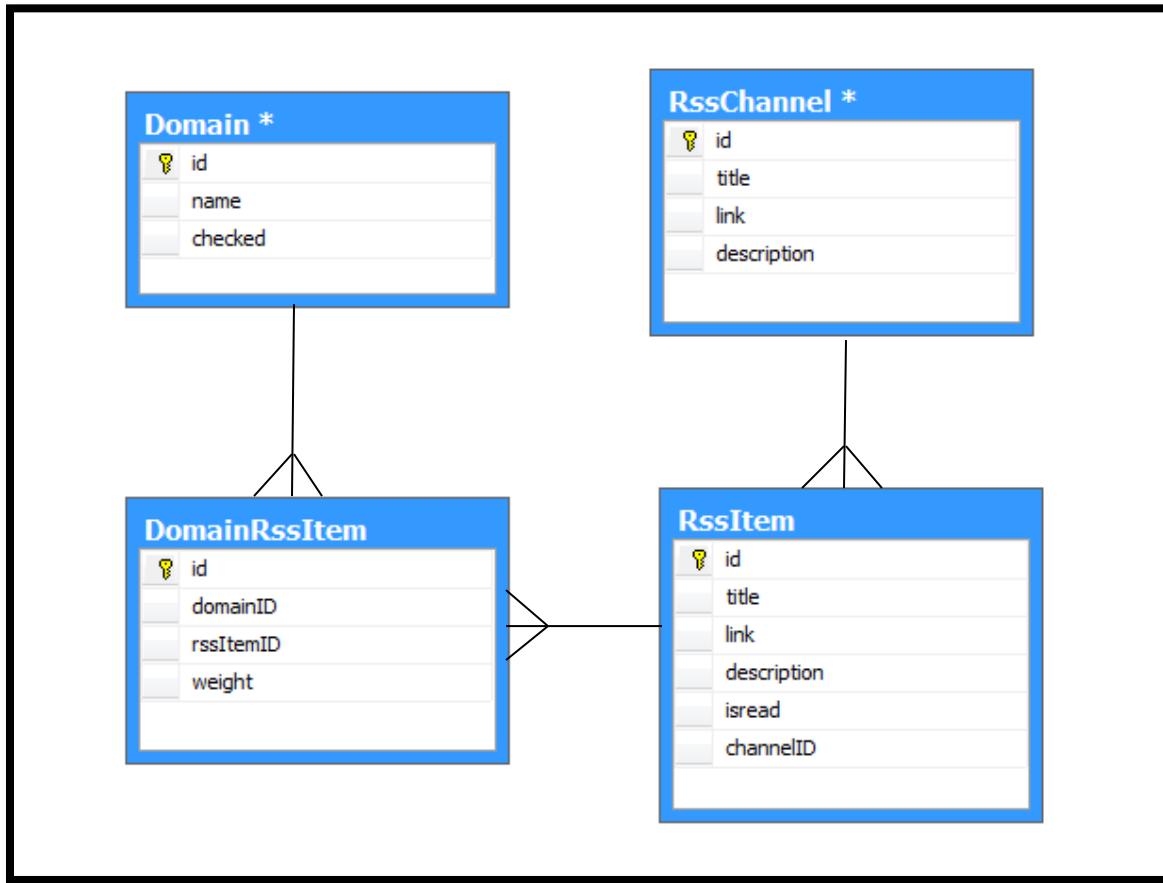
## 2 العلاقات بين الكيانات

1- هناك علاقة واحد لكثير بين الكيان RssChannel والكيان RssItem ، فمن أجل كل قناة أخبار هناك

مجموعة من الأخبار الموجودة ضمنها ؛

2- هناك علاقة كثير لكثير بين الجدول Domain والجدول RssItem ، فالمجال الواحد له أكثر من خبر والخبر

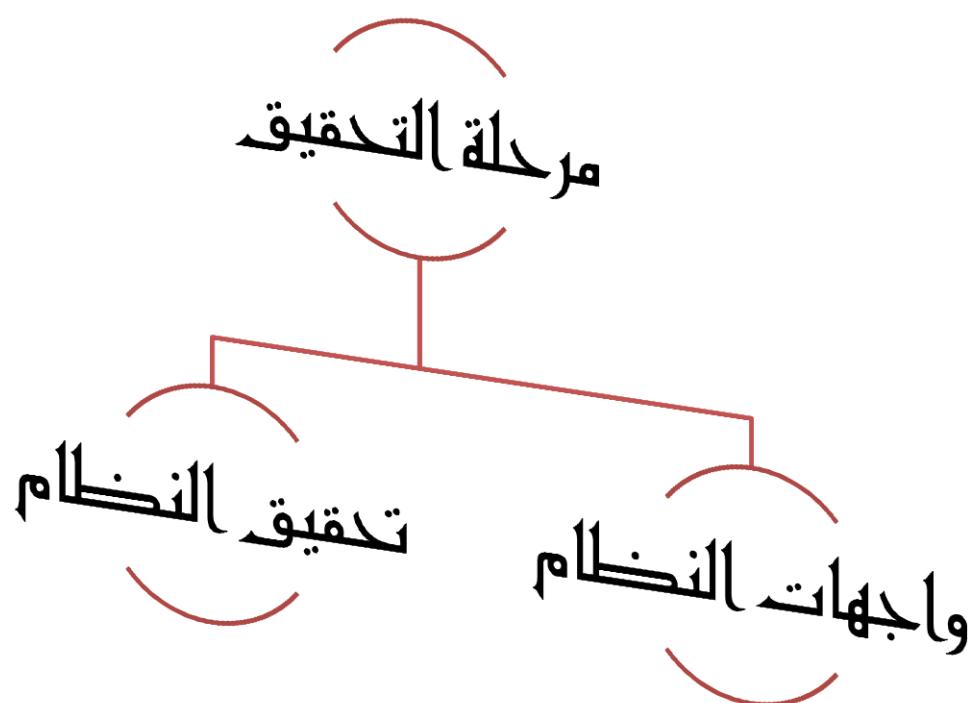
.DomainRssItem والواحد يتبع لأكثر من مجال ، تم كسر هذه العلاقة من خلال الجدول



الشكل 56 مخطط قاعدة المطبيات

## الباب الخامس

### مرحلة التحقيق



الشكل 7 5 النقاط الأساسية ضمن مرحلة التحقيق



## الفصل الثاني عشر

### تدقيق النّظام

- قمنا ببناء خدمة الويب web service التي تعمل على الفهم الدلالي لمحظى الصفحة، ومن ثم تم الاستفادة من

هذه الخدمة في ثلاث تطبيقات كما ذكرنا سابقا.

■ مخدم البروكسي؛

■ القارئ؛

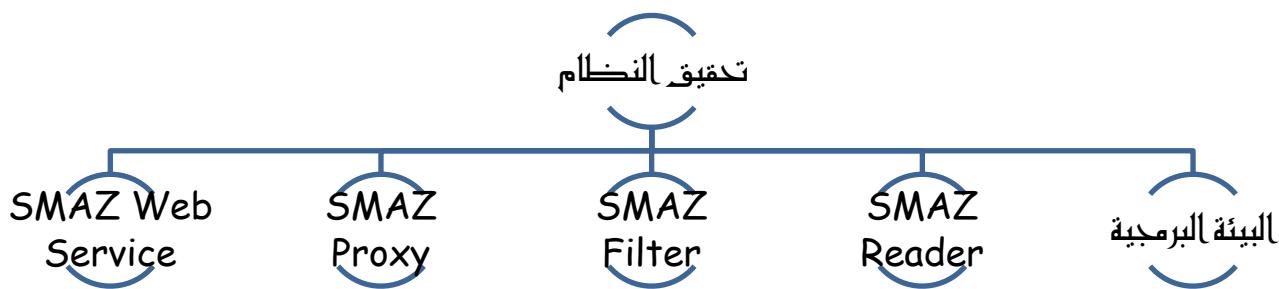
■ إضافة المتصفح.

- يتم وضع إعدادات التطبيق في ملف الإعدادات؛

- من أجل التطبيقيين مخدم البروكسي وإضافة المتصفح هناك قواعد يتم كتابتها ووضعها في ملف القواعد من خلال

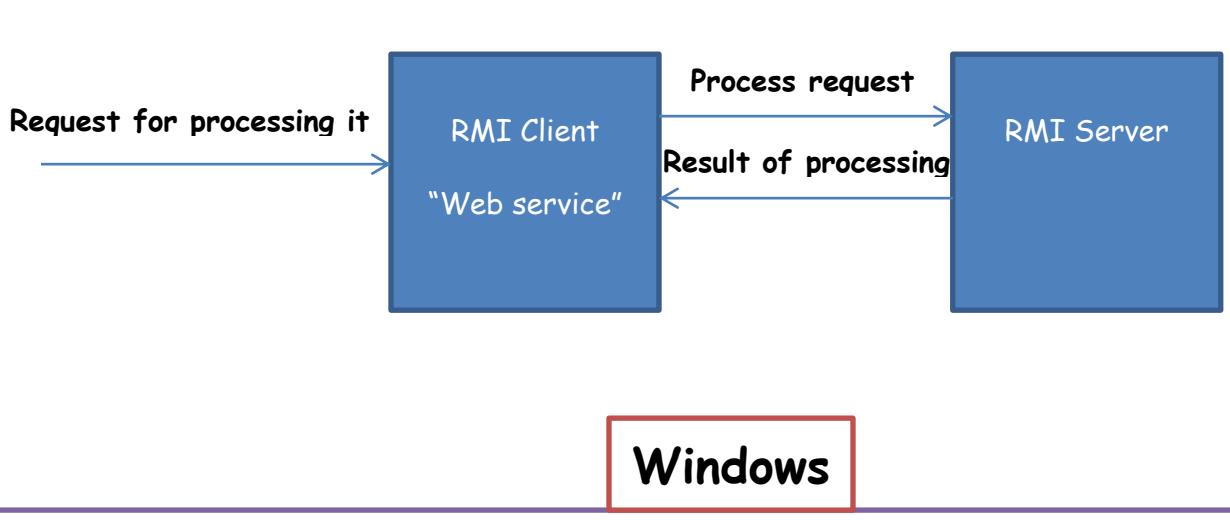
واجهة تساعد على ذلك، يتم الاستفادة من هذه القواعد للإقرار بحجب الصفحة أم لا؛

- في تطبيق الموبايل "القارئ" لا تحتاج إلى قواعد، لأن ما نقوم به هو تصنيف للأخبار وليس فلترة.



الشكل 58 النقاط الأساسية في تحقيق النظام

## فكرة RMI 1.1



الشكل 59 مخطط يوضح آلية عمل خدمة الويب

- Remote Method Invocation ،

- واجهتنا مشكلة أثناء العمل وهي أننا من أجل كل طلب لصفحة انترنت سيتم إعادة تنفيذ الخطوات من البداية بما

فيها تحميل المكتبات التي تحتاج إليها، تحميل هذه المكتبات يأخذ وقتاً طويلاً وما يهمنا فعلاً هو تقليل زمن تنفيذ

العملية قدر الإمكان كونها تتم online؛

- من هذه الحاجة ظهرت لنا فكرة الاستفادة من تقنية RMI، حيث يكون لدينا مخدم rmi srever ويكون

- لدينا زبون rmi client، هذا المخدم مسؤول عن عمليات تحميل المكتبات والملفات اللاحزة وتبقى محملة في

الذاكرة طيلة الوقت، يبقى المخدم في حلقة مستمرة دائماً ينتظر ورود الطلبات إليه، فمن أجل كل طلب يقوم الـ

ـ icap server ي تقوم هذا الأخير بطلب خدمة web service والتي تمثل بحد ذاتها الـ rmi client

طلب تنفيذ التابع process request الموجود عند المخدم rmi server، وتنفيذ التابع عند المخدم ورد

النتيجة إلى الـ rmi client ومن ثم إلى icap server

- لتحقيق هذه التقنية تحتاج بداية إلى remote interface هي SemanticAnalyzer extends

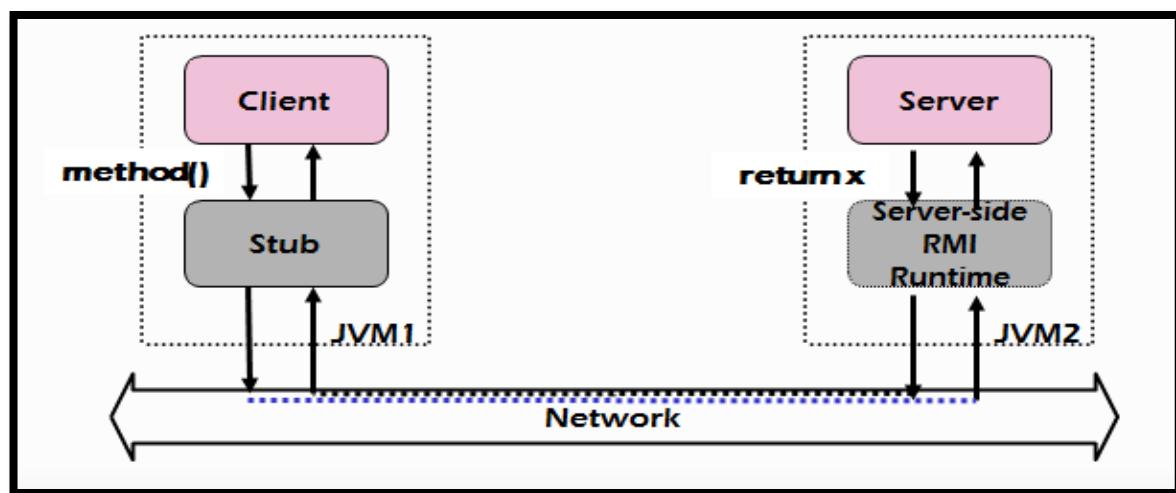
نعرف فيها التابع: Remote

- 1- load المسئول عن تحميل المكتبات اللازمة ويرد true في حال كانت كل المكتبات قد تم تحميلها بشكل

صحيح؛

- 2- processRequest المسئول عن معالجة الطلبات ويرد سلسلة محرافية تحوي كل مجال domain ونسبة وروده في النص.

ونحتاج إلى Rmi client و Rmi server



الشكل 60 آلية عمل RMI

- ومن الجدير بالذكر أن تقنية الـ RMI تختفي تفاصيل الاتصالات بين المخدم والزبون.

- تقوم ببناء rmi server والذي يقوم بعمل تحقيق للواجهة السابقة وإعادة تعريف للتتابع الموجودة فيها؛
- ضمن ال rmi server تقوم بعمل نسخة من غرض التنفيذ وتسجيله ضمن rmi registry وذلك من خلال

التعليمية :

```
Registry registry = LocateRegistry.createRegistry(port); 1
SemanticAnalyzer sa = new SemanticAnalyzerImpl(); 2
registry.rebind("sa", sa); 3
```

- بناء rmi registry ليتم من خلالها تحويل غرض التنفيذ إلى غرض شبكي وحفظه ضمنها ليقوم الزبون

لاحقاً بالبحث عن هذا الغرض ضمنها والحصول على مؤشر عليه؛

- عمل نسخة من غرض التنفيذ؛
- تسجيل غرض التنفيذ ضمن rmi registry للحصول عليه لاحقاً من قبل الزبون والاستفادة من عملياته
- التي يقدمها.

- ضمن ال rmi client يتم الحصول على نسخة من غرض التنفيذ بعد البحث عليه ضمن rmi registry واستدعاء إحدى عملياته وهي processRequest من أجل صفحة الإنترنت المطلوبة ومعالجتها عند المخدم وإرسال النتيجة النهائية للزبون

```
Registry registry = LocateRegistry.getRegistry(serverIP, port);
```

1

```
SemanticAnalyzer sa = (SemanticAnalyzer)registry.lookup("sa");
```

2

1- يتم الحصول على ال rmi registry التي تم بناؤها سابقاً؛

2- يتم البحث عن غرض التنفيذ ضمن ال rmi registry والذي يحمل الاسم sa والحصول على مؤشر على غرض

التنفيذ؛

3- بعد ذلك يصبح rmi client يملك مؤشر على غرض التنفيذ وبإمكانه استدعاء أي خدمة يقدمها rmi server مثل

معالجة الطلبات .sa.processRequest

## 2 بنيـة ملف الإـعدادات Config File

- يحـوي هذا المـلف مـجمـوعـة من الـبارـامـترـات المـحدـدة مـسـبـقاً وـيـكـون مـوـجـود عـنـدـ الـIcap~Server وـهـو بـدورـه

client لـلـweb~service المسـؤـولـة عـن حـسـاب نـسـبـ الـdomains المـوـجـودـة فـي صـفـحةـ الـاـنـتـرـنـتـ المـرـرـةـ أوـ

الـملـفـ المـرـرـ للـخـدـمـةـ؛

- هـذه الـبارـامـترـات هـيـ:

▪ serviceURL عنـوان خـدـمـةـ الـw~i~pـ الـتـي سـيـتـم الـاـتـصـال بـهـا وـالـاستـفـادـةـ مـنـ خـدـمـاتـهـاـ؛

▪ getDocContentFromProxy متتحول منطقي يأخذ قيمة true عندما نريد من أن يقوم بتحميل محتوى الصفحة أو الملف بنفسه و false عندما يقوم icap server بإرسال المحتوى إليه

▪ sortDocDomainsByWeight متتحول منطقي يأخذ قيمة true عندما نريد ترتيب الصفحة تنازلياً حسب نسبة وردوهم ضمن الصفحة؛ domains

▪ addWeightsToAncestors متتحول منطقي يأخذ قيمة true عندما نريد رفع وزن الكلمة التابعة ل domain ما إلى آباء هذا ال domain فعادة عند ورود كلمة تنتهي إلى domain ما بنسبة 0.5 فنعطي حرية رفع هذا الوزن إلى آباء هذا ال domain أم لا من خلال هذا المتتحول المنطقي؛

▪ shortWordsLimit متتحول صحيح يدل على الطول الأصغرى للكلمة التي يجب أخذها في الحساب فإذا كانت قيمته مثلاً 3 فعندها كل كلمة طولها أصغر من 3 لن يتم أخذها بعين الاعتبار عند المعالجة؛

▪ searchInDbpedia متتحول منطقي يأخذ قيمة true عندما نريد البحث عن الكلمة غير موجودة ضمن ال wordnet ضمن ال dbpedia ففي حال لم تكن الكلمة موجودة ضمن ال wordnet وكان هذا المتتحول المنطقي true عند سистем البحث عن هذه الكلمة ضمن ال dbpedia، dbpedia

▪ dbpediaResultsLimit متتحول صحيح يدل على عدد القيم التي سيتم أخذها بعين الاعتبار من أجل كل علاقة سيتم الاستفادة منها من ال dbpedia، بمعنى أنه عندما لا نجد الكلمة ضمن ال wordnet سيتم البحث عنها ضمن ال dbpedia، صفحة ال dbpedia تحوي مجموعة من العلاقات التي من الممكن الاستفادة منها في محاولة لفهم معنى الكلمة، من هذه العلاقات هي علاقة subject، فمن أجل هذه العلاقة ومن أجل الكلمة المطلوبة هناك مجموعة من القيم التي تشرح معنى هذه الكلمة وبالتالي لابد من تحديد العدد الأعظمي للقيم التي سيتم أخذها بعين الاعتبار؛

disambiguationMethod متحول صحيح يدل على الخوارزمية المتبعة لإزالة الغموض فكما ذكرنا ▪

سابقاً كنا قد حققنا مجموعة من الخوارزميات لإزالة غموض كلمة كل منها مختلفة في تعقيدها وسرعتها ودققتها

عن الأخرى

○ الرقم 1 يعني أن الخوارزمية المتبعة هي خوارزمية المعنى الأكثر شهرة؛

WordNet Semantically Tagged ○ الرقم 2 يعني أن الخوارزمية المتبعة هي خوارزمية

,glosses

○ الرقم 3 يعني أن الخوارزمية المتبعة هي خوارزمية Hood؛

○ الرقم 4 يعني أن الخوارزمية المتبعة هي خوارزمية Lesk.

acceptedTags للدلالة على نوع الكلمات التي نرغب بمعالجتها ضمن النص سواء كانت فعل، اسم، ▪

صفة أو ظرف؛

accepted tags

NN=noun

VB=verb

JJ=adjective

RB=adverb

withCoreference متحول منطقي يأخذ قيمة true عندما نريد إعادة الفضائل إلى أصلها أثناء معالجة ▪

النص المطلوب؛

dbpediaPredicates لتحديد العلاقات التي نريد معالجتها والبحث فيها ضمن أنطولوجية ▪

dbpedia searchInDbpedia عندما يكون خيار dbpedia مفعول؛

### 3 بنية ملف القواعد Rules File

- يحتوي هذا الملف على مجموعة القواعد المجموعة من قبل خبراء بهدف حجب الصفحات بناء على محددات

وغایات خاصة؛

- بعد وضع القواعد من قبل خبراء وباستخدام واجهة مساعدة يتم حفظ هذه القواعد في ملف القواعد الموجود عند الـ

Icap server والذي يقارن بين القواعد المجموعة وبين نتائج معالجة النص المطلوب لمعرفة فيما إذا كانت

الصفحة أم الملف سيحجب أم لا؛

- تتكون القاعدة في الحالة العامة من ثلاثة أقسام:

○ القسم الأول: مجموعة من الـ domains مع نسبة وجود كل domain ضمن الصفحة؛

○ القسم الثاني: مجموعة من الكلمات تفصل بينها علامة (&) ويتبع الـ designer للخبير وضع كلمة

مع معانيها بحيث نفصل بين كل كلمة ومعناها بفاصلة (,) وتعني سواء وردت هذه الكلمة أو أي كلمة

من مرادفاتها؛

○ القسم الثالث الحدث Action الذي يجب تطبيقه عند تحقق القاعدة وقد يكون حجب الصفحة أو

تمريرها، B,U

- مثال:

Sexuality >= 0.2 & psychoanalysis >= 0.1 (sex, porn & girl) : u;

$\text{Sexuality} \geq 0.2 : b;$

- القاعدة الأولى يقصد بها عند ورود المجال Sexuality بنسبة أكبر أو تساوي 0.2 والمجال psychoanalysis بنسبة أكبر أو تساوي 0.1 يكون قد تحقق القسم الأول للقاعدة، عند تحقق القسم الأول وورود الكلمة sex أو مرادفتها girl مع الكلمة porn ستكون قد تحققت القاعدة كاملة عندها فقط يتم تنفيذ القاعدة الأولى وهو عدم حجب الصفحة U، تكون القاعدة قد تحققت
- القاعدة الثانية يقصد بها عند ورود المجال Sexuality بنسبة أكبر أو تساوي 0.2 ت تكون القاعدة قد تحققت وال action هو حجب الصفحة B،
- ترتيب القواعد مهم جدا ففي حال كانت القاعدة غير محققة يتم الانتقال إلى القاعدة التي تليها حتما، وفي حال كانت القاعدة محققة والحدث هو عدم حجب الصفحة أو حجب الصفحة لا يتم الانتقال إلى القاعدة التي تليها،
- العمليات التي من الممكن تطبيقها على المجالات هي <,>, =, <=, =>;
- قمنا ببناء rule editor لبناء القواعد ومن ثم تخزينها في ملف القواعد،
- هناك مفسر للقواعد rule parser يعمل على قراءة القواعد وتحليلها وفهمها.

### Icap server 3.1

- هو زبون لخدمة الويب client rmi أي أن:

Icap server = client for web service (rmi client)

- عند طلب صفحة الويب من قبل الزبون يمر الطلب من خلال مخدم squid proxy server، ليتم تحرير

الطلب بعدها إلى icap server الذي يحوي كود استدعاء خدمة الويب بعد الحصول على بارامترات الاستدعاء

من خلال ملف الإعدادات الموجود عنده config file؛

- يقوم icap server بالاتصال بال rmi server ، ليقوم هذا الأخير بفهم النص دلائيا وتنفيذ تابع معالجة

الطلب الذي يقدمه والحصول على كل domain مع نسبة وجوده ضمن النص ، تُرد هذه النتيجة إلى icap

server الذي يقوم بمقارنة نتائج الطلب مع القواعد الموجودة عنده ضمن ملف القواعد rules file بالترتيب؛

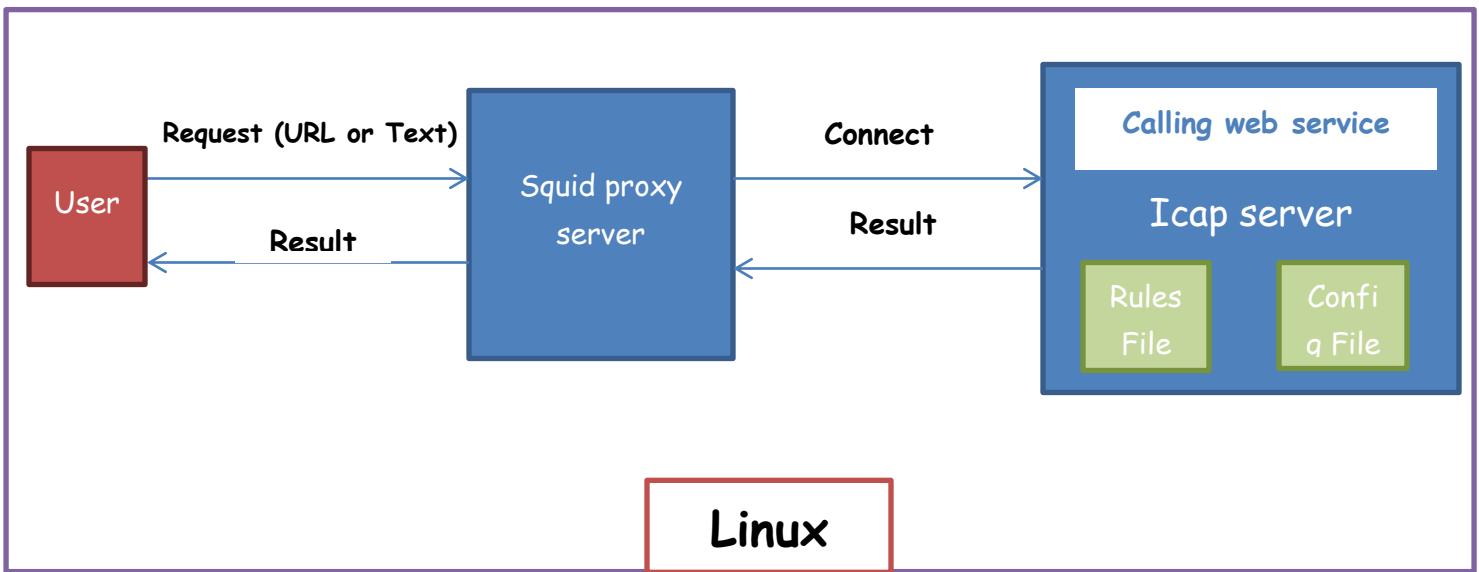
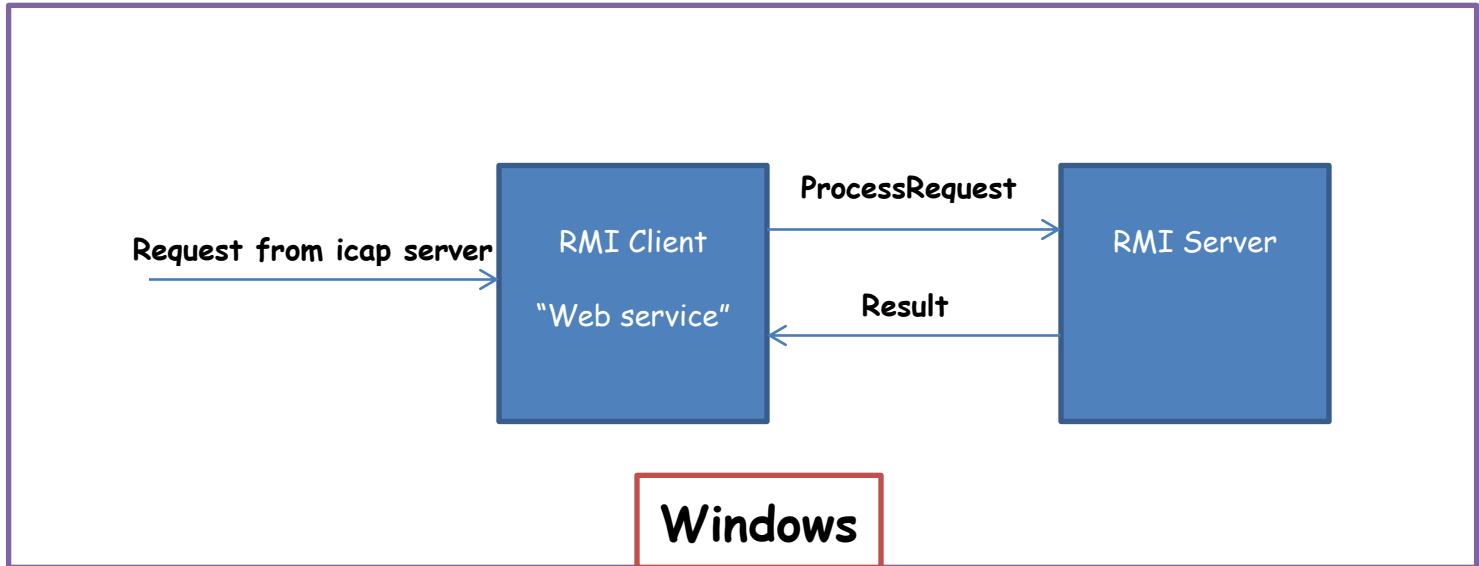
- اعتماداً على هذه المقارنات يتم الحصول على نتائج الطلب هل يتم حجبه أم لا، في حال كان هناك مطابقة مع

قاعدة تقر بعد الحجب يتم تمرير الصفحة وعرضها للزبون ويتم التوقف عن مسح بقية القواعد، وفي حال كان هناك

مطابقة مع قاعدة تقر بحجب الصفحة يتم حجب الصفحة وعدم عرضها على الزبون والتوقف عن مسح بقية

القواعد.

### 3.2 مخطط العمل وآلية الاتصال بين مكونات النظام



الشكل 61 مخطط العمل وآلية الاتصال بين المكونات

- هنا يتم العمل ضمن المتصفح مباشرة من دون المرور بمخدم البروكسي squid و icap ،
- عند طلب الصفحة من خلال المتصفح أو تحديث الصفحة يتم معالجة الطلب من خلال خدمة الويب التي تعمل على فهم الصفحة ورد النتيجة ، ليتم بعدهاأخذ قرار حجب الصفحة أم لا من خلال القواعد المحددة مسبقاً ،
- في حال كانت القرار هو حجب الصفحة يتم عرض الصفحة .errorPage.html



الشكل 62 صفحة الحجب

- تطبيق موبайл يعمل على تصنيف الأخبار إلى مجالات محددة وفقاً للمحتوى؛
- يتم إضافة قناة أخبار من خلال تحديد الرابط الخاص بها؛
- يتم الحصول على الأخبار الموجودة ضمن الصفحة، وطلب خدمة الويب ومعالجة كل عناصر الأخبار؛
- نتيجة الاستعانة بهذه الخدمة يتم تصنيف الأخبار إلى مجالات معينة.

### اختبار أداء Smaz Proxy

- لكي نتمكن من تحديد مدى فاعليه Smaz Proxy و قابليته للاستخدام الفعلى، قمنا بإجراء بعض الاختبارات التي تبيّن مدة تحميل بعض صفحات الانترنت مع استعمال Smaz Proxy وبدون استعماله؛
  - قمنا بإجراء الاختبارات باستعمال حاسب (مخدم عتادي) يحمل المواصفات التالية:
    - 1- معالج من نوع Intel Core I7 طراز 2.2 GHz بتردد 4 GHz
    - 2- ذاكرة عشوائية بسعة 4 GB.
- وباستعمال اتصال انترنت ADSL 1 mbps، من دون تفعيل خيار Mozilla Firefox 16، و متصفح coreference، وباستخدام الخوارزمية DBpedia البحث في WordNet tagged لـ *glosses* لإزالة الغموض

فحصلنا على النتائج التالية :

الصفحة	زمن 1	زمن 2
www.huffingtonpost.com/news/syria	13.64	29.30
www.usnews.com/photos/violence-in-syria	54.77	61.43
www.insectidentification.org/cockroaches.asp	13.65	23.54
en.wikipedia.org/wiki/Politics_of_Israel	22.66	26.12
en.wikipedia.org/wiki/Sex	25.34	48.62
www.olympic.org/swimming-100m-backstroke-women	25.47	54.59

### جدول 27 اختبار أداء SMAZ Proxy

حيث :

- زمن 1 : هو زمن تحميل الصفحة باستعمال مخدم proxy بدون اي اضافات ؛
- زمن 2 : هو زمن تحميل الصفحة باستعمال مخدم Smaz Proxy الذي يعتمد على الفلترة الدلالية ؛
- وجميع الازمنة مقدرة بالثانية ؛

## نتيجة

بعد قيامنا بتجربة عدد لا يأس به من الصفحات و تحليل النتائج، لاحظنا ان استعمال Smaz Proxy يسبب زيادة حوالي 70٪ في مدة تحميل الصفحة و هذه نسبة مقبولة جدا نظرا للخدمات التي يقدمها.

## ملاحظة

ضمن بيئه الاختبار الخاصة بنا تم تثبيت ال Ubuntu Linux Proxy Server على نظام تشغيل Ubuntu Linux ضمن آلة افتراضية VirtualMachine وهذا أدى الى بطء في الأداء إلى حد ما، لكن عند يتم ثبيت ال Proxy على جهاز مستقل و بنظام تشغيل حقيقي سنلاحظ تحسن ملحوظ في السرعة.

## 6 البيئة البرمجية

### 6.1 لغة البرمجة المستخدمة

يعد اختيار لغة البرمجة المناسبة مرحلة مهمة، خاصة في الأنظمة المعقّدة التي تتطلب شروطاً خاصة بها مثل وجود عدة أنظمة تشغيل أو عدة أنواع من نظم إدارة قواعد المعطيات، كل هذا وغيرها يجعل اختيار لغة البرمجة المناسبة مرحلة مهمة من المستحق الوقوف عندها.

إن لغة الجافا هي لغة مناسبة جداً لهذا النظام لما تتمتع به من ميزات، أهمها الاستقلالية عن نظام التشغيل وعدة أسلوب آخرى منها:

- لغة برمجة غرضية التوجه بشكل كامل؛

- المحمولية؛

- المقرؤية وتنظيم اللغة؛
- القياسية؛
- الصيانية وسرعة التطوير؛
- أمن اللغة؛
- معالجة الاستثناءات والأخطاء؛
- دعم النياسب المتعددة؛
- دعم واجهة المستخدم البيانية؛
- الانتشار والتسويق والدعم الفني؛
- يوجد لها الكثير من المصادر على شبكة الإنترنت.

Java -

تم استخدام لغة البرمجة جافا في بناء التطبيقات الثلاثة الأساسية :

- Rmi server
  - Rmi client
  - Icap server
- وذلك باستخدام eclipse enterprise edition و netbeans .
- تم برمجة تطبيق الموبايل من خلال Java for android ،
- في تطبيق المتصفح تم استخدام JavaScript و HTML و CSS و ملف توصيف لهذه الإضافة وهو مكتوب بلغة JSON ،
- في تطبيق الموبايل تم استخدام SQLite للتعامل مع قواعد المعطيات.

## jsp web service 6.2

- تأمين خدمة الويب بشكل سهل.

## ICAP 6.3

Internet Content Adaption Protocol -

GreasySpoon -

- يتيح هذا البروتوكول للمستخدم icap client بتمرير رسالة إلى مخدم icap server لإجراء عمليات المعالجة

[11] "adaptation"

- يقوم المخدم بتنفيذ عمليات التحويل والمعالجة المطلوبة وإرسال النتيجة إلى الزبون، عادة ما تكون الرسالة معدلة؛

- غالباً ما تكون الرسالة هي طلب http request أو استجابة http response

- عادة يستخدم لتحقيق مضاد فيروسات virus scanning .content filters Icap -

## 6.4 المكتبات المستخدمة

### Apache Jena 6.4.1

- مكتبة مستخدمة في الجافا من أجل بناء تطبيقات الويب الدلالي؛

- Jena تقدم مجموعة من الأدوات ومكتبات الجافا التي تساعد في تطوير تطبيقات الويب الدلالي والمعطيات

المترابطة Linked data

Jena تتضمن:

○ API لقراءة ومعالجة وكتابة ثلاثيات الـ RDF في ملفات XML ،

○ القدرة على بناء أنطولوجيات باستخدام OWL والتعامل مع أنطولوجيات موجودة بصيغة RDF؛

- محرك بحث مزود بأخر توصيف للغة الاستعلام Sparql،
- تخزين عدد كبير من ثلاثيات ال RDF ليتم استخدامها بشكل فعال لاحقاً،
- إمكانية نشر ثلاثيات ال RDF إلى تطبيقات أخرى باستخدام مجموعة من البروتوكولات بما فيها sparql،
- محرك استدلال قائم على قواعد التفكير مع مصادر البيانات RDF و OWL.
- تشبه مكتبة ال SemWeb و dotnetrdf الموجوده في ال C#.

#### [Apache PDFBox 6.4.2](#)

- أداة في الجافا مفتوحة المصدر تستخدم للتعامل مع مستندات ال PDF؛
- تتيح بناء مستندات PDF والتعامل مع مستندات موجودة مسبقاً والقدرة على استخراج محتواها؛
- تم استخدام هذه المكتبة في المشروع لاستخراج محتوى مستندات ال PDF والحصول على النص الموجود فيها.

#### [Apache POI 6.4.3](#)

- أداة في الجافا مفتوحة المصدر تستخدم للتعامل مع مستندات ال office؛
- تتيح بناء مستندات office والتعامل مع مستندات موجودة مسبقاً والقدرة على استخراج محتواها؛
- تم استخدام هذه المكتبة في المشروع لاستخراج محتوى مستندات ال word ولا سيما ملفات word والحصول على النص الموجود فيها.

#### [Boilerpipe 6.4.4](#)

- أداة في الجافا مفتوحة المصدر تستخدم للتعامل مع صفحات ال h1tm،
- تتيح بناء صفحات html والتعامل مع صفحات موجودة مسبقاً والقدرة على استخراج محتواها؛

- تم استخدام هذه المكتبة في المشروع لاستخراج محتوى صفحات ال html والحصول على النص الموجود فيها.

#### Jsoup 6.4.5

- المكتبة تسمح بالتعامل مع صفحات HTML، حيث أنها تقوم بتحويل صفحة HTML إلى مجموعة من العقد

والتي تمكن من البحث واستخراج المعلومات بسهولة، واستخراج العقد التي نريد بسرعة، كما تستطيع إنشاء

صفحات أو إضافة عقد إليها أو تغيير عقد موجودة؛

- هذه المكتبة مشابهة HtmlAgilityPack لكنها تستخدم في الجافا، تم استخدام هذه المكتبة للحصول على

النص المرافق للصور في صفحات ال html ، وذلك بهدف معالجة النص المحيط بالصور بغية حجب الصفحات

التي تحوي صور يجب حجبها؛

- مثال عن استخدام هذه المكتبة حيث يتم الحصول على صفحة html ومن ثم الحصول على عقد محددة

```
Document doc = Jsoup.connect("http://en.wikipedia.org/").get();
Elements newsHeadlines = doc.select("#mp-itn b a");
```

#### JWI 6.4.6

Java Wordnet Interface -

- واجهة في الجافا تؤمن عمليات التعامل مع معجم ال wordnet؛

- تؤمن عمليات الوصول إلى المجموعات والتعامل معها بشكل مباشر والحصول على المجموعات synset ومعنى كل

مجموعة gloss ومجموعة المترادفات وغير ذلك من العمليات؛

- تؤمن عمليات الحصول على كلمة معينة والحصول على الجذر المجرد الخاص بها lemma ومعناها بشكل سهل

ومباشر؛

- مثال عن هذه العمليات:

```
public void runExample(){  
  
    // construct the URL to the Wordnet dictionary directory  
    String wnhome = System.getenv("WNHOME");  
  
    String path = wnhome + File.separator + "dict";  
    URL url = null;  
    try{ url = new URL("file", null, path); }  
    catch(MalformedURLException e){ e.printStackTrace(); }  
    if(url == null) return;  
  
    // construct the dictionary object and open it  
    IDictionary dict = new Dictionary(url);  
    dict.open();  
  
    // look up first sense of the word "dog"  
    IIndexWord idxWord = dict.getIndexWord("dog", POS.NOUN);  
    IWordID wordID = idxWord.getWordIDs().get(0);  
    IWord word = dict.getWord(wordID);  
    System.out.println("Id = " + wordID);  
    System.out.println("Lemma = " + word.getLemma());  
    System.out.println("Gloss = " + word.getSynset().getGloss());  
}
```



1



2



3

1- الحصول على مكان وجود معجم wordnet ، المسار الموجود ضمنه المعجم؛

2- الحصول على غرض من المعجم ليتم التعامل معه؛

3- مجموعة من العمليات توضح آلية التعامل مع المعجم والذي تتيحه مكتبة jwi

a. الحصول على أول معنى لكلمة “dog”؛

b. الحصول على مميز هذه الكلمة؛

c. الحصول على الكلمة؛

d. الحصول على جذر الكلمة؛

e. الحصول على معنى المجموعة الموجودة ضمنها كلمة “dog”.

#### Stanford CoreNLP 6.4.7

- مكتبة تؤمن لنا القدرة على إجراء عمليات معالجة اللغات الطبيعية :

○ تقسيم النص إلى مجموعة جمل وتدعى هذه العملية splitting؛

○ تقسيم الجملة إلى مجموعة كلمات وتدعى هذه العملية tokenizing؛

○ رد كل كلمة إلى أصلها الموجود ضمن معجم wordnet وتدعى هذه العملية lemmatizing؛

○ معرفة موقع كل كلمة ضمن النص هل هي فعل، فاعل، صفة أو ظرف وتدعى هذه العملية pos

؛tagging

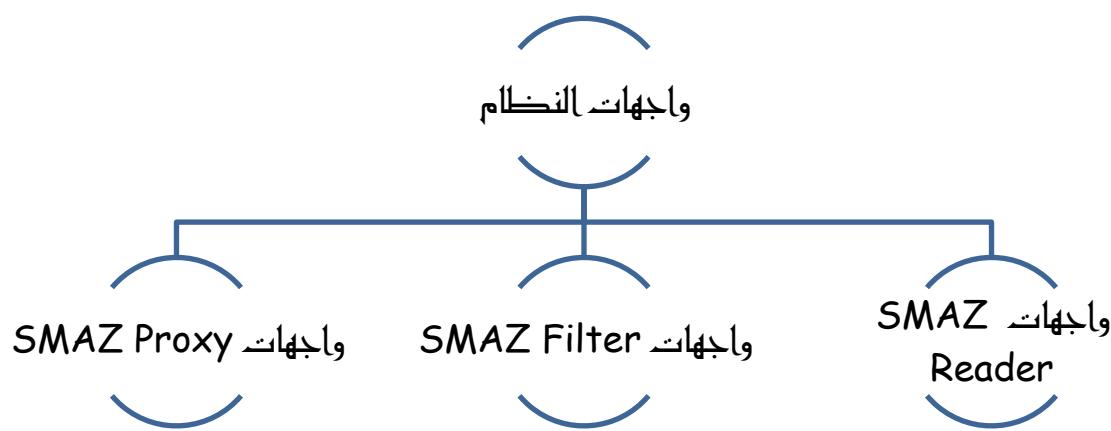
○ رد الضمائر إلى أصلها وتدعى هذه العملية coreferenceing.

- Spotlight هي أداة للتأشير التلقائي للنصوص على موارد ال DBpedia؛
- توفير حل لربط مصادر المعلومات غير المهيكلة لفتح سحابة البيانات المرتبطة من خلال DBpedia؛
- توفر web service دخلها النص وخرجها ال annotated text؛
- يمكن تمرير النص كامل وتحديد كلمات محددة للتأشير عليها؛
- قمنا بتمرير الجملة كاملة في محاولة لإزالة غموض الكلمات، حيث تقدم ال spotlight أيضا خدمة فك غموض النص الممرر لها. [10]



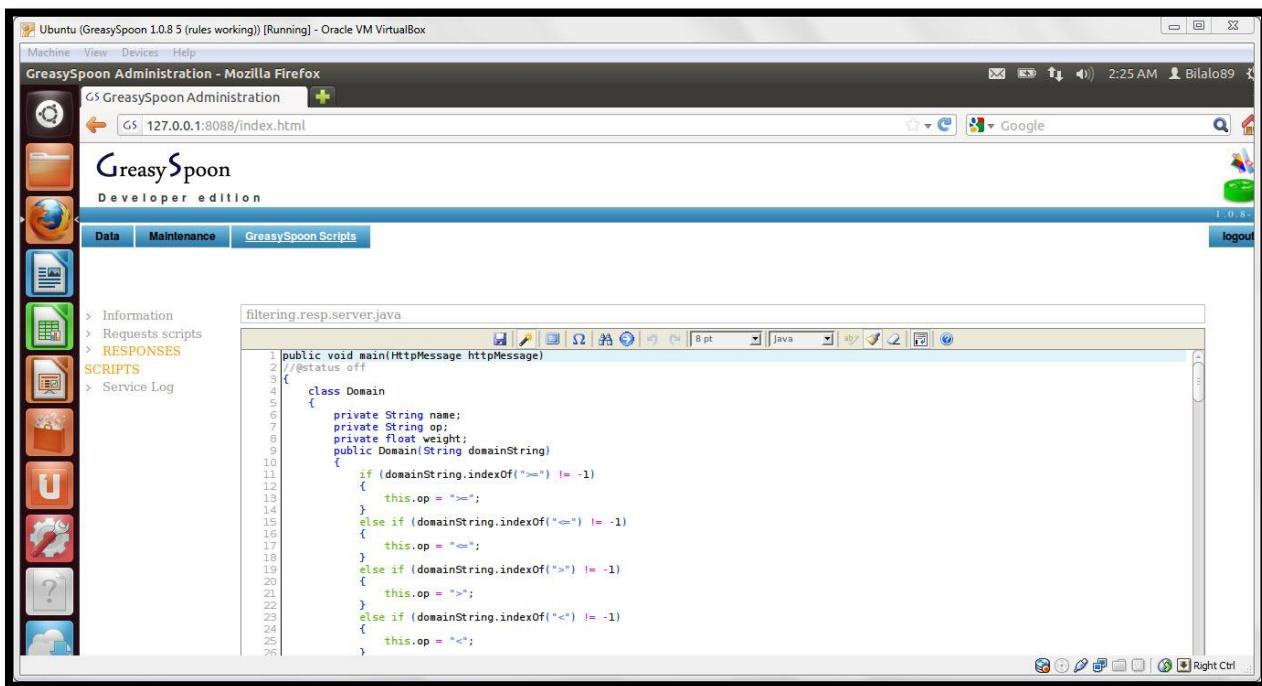
## الفصل الثالث عشر

### واجهات النظام



الشكل 63 النقاط الأساسية ضمن واجهات النظام

## 1.1 مخدم icap والبرمجة بداخله



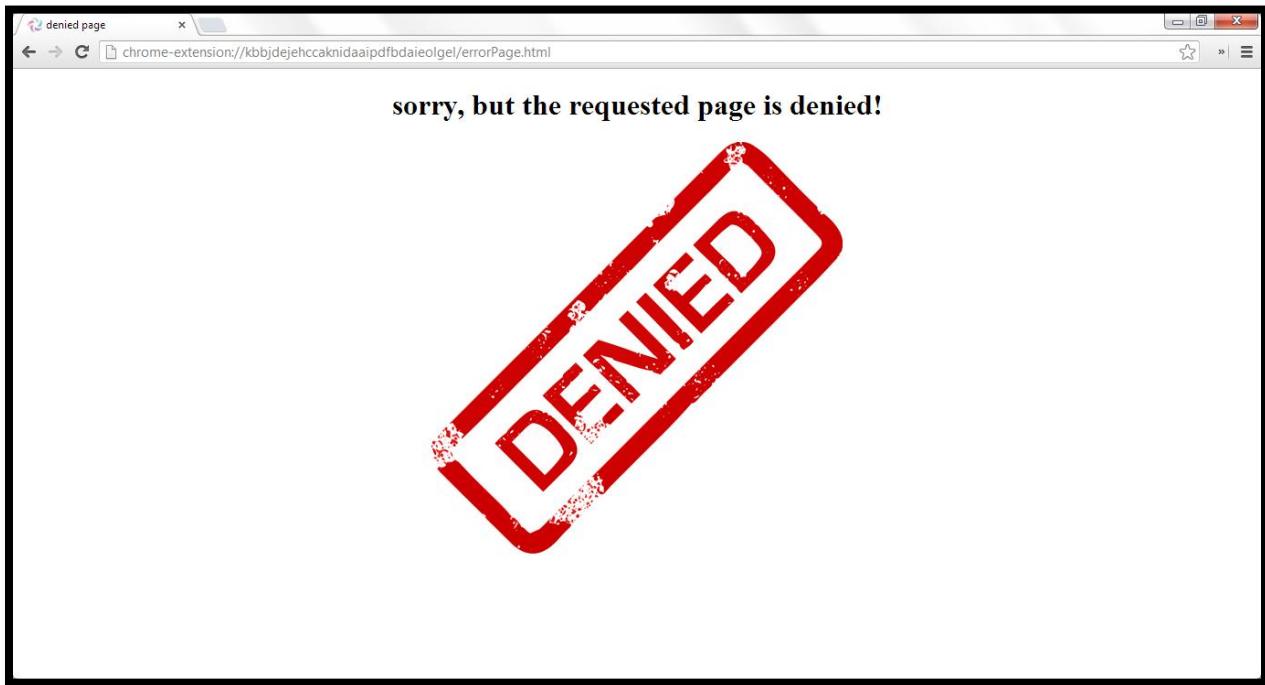
الشكل 64 مخدم icap والبرمجة بداخله

## 1.2 التأكيد من أن الخدم متصل وقدر على معالجة الطلبات

```
04. Command Prompt - java -jar SemanticAnalyzer.jar
loading domain 154: telecommunication.ppv
loading domain 155: telegraphy.ppv
loading domain 156: telephony.ppv
loading domain 157: tennis.ppv
loading domain 158: theatre.ppv
loading domain 159: theology.ppv
loading domain 160: time_period.ppv
loading domain 161: topography.ppv
loading domain 162: tourism.ppv
loading domain 163: town_planning.ppv
loading domain 164: transport.ppv
loading domain 165: tv.ppv
loading domain 166: university.ppv
loading domain 167: vehicles.ppv
loading domain 168: veterinary.ppv
loading domain 169: volleyball.ppv
loading domain 170: wrestling.ppv
-----
loading Gloss Sysnsets WSD...
-----
loading Jena...
-----
Server is connected and ready to process requests.
```

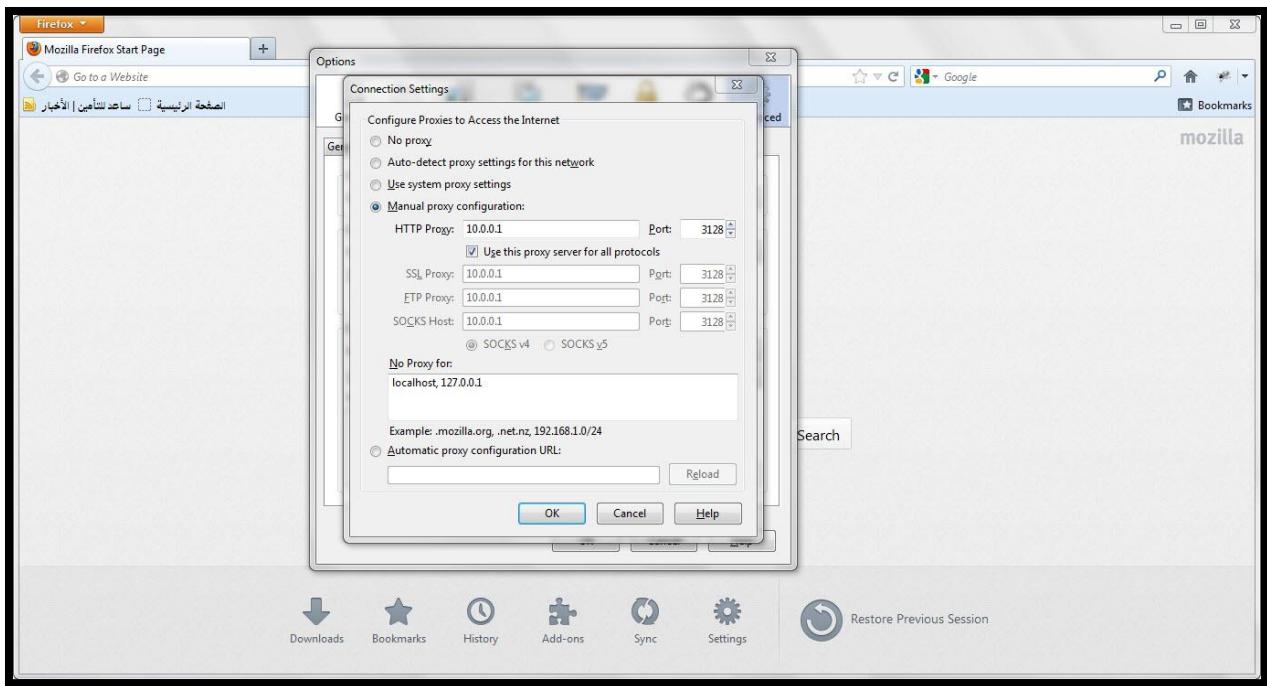
الشكل 65 التأكيد من أن المخدم متصل وقدر على معالجة الطلبات

### 1.3 طلب صفحة ومعالجتها وحجبها



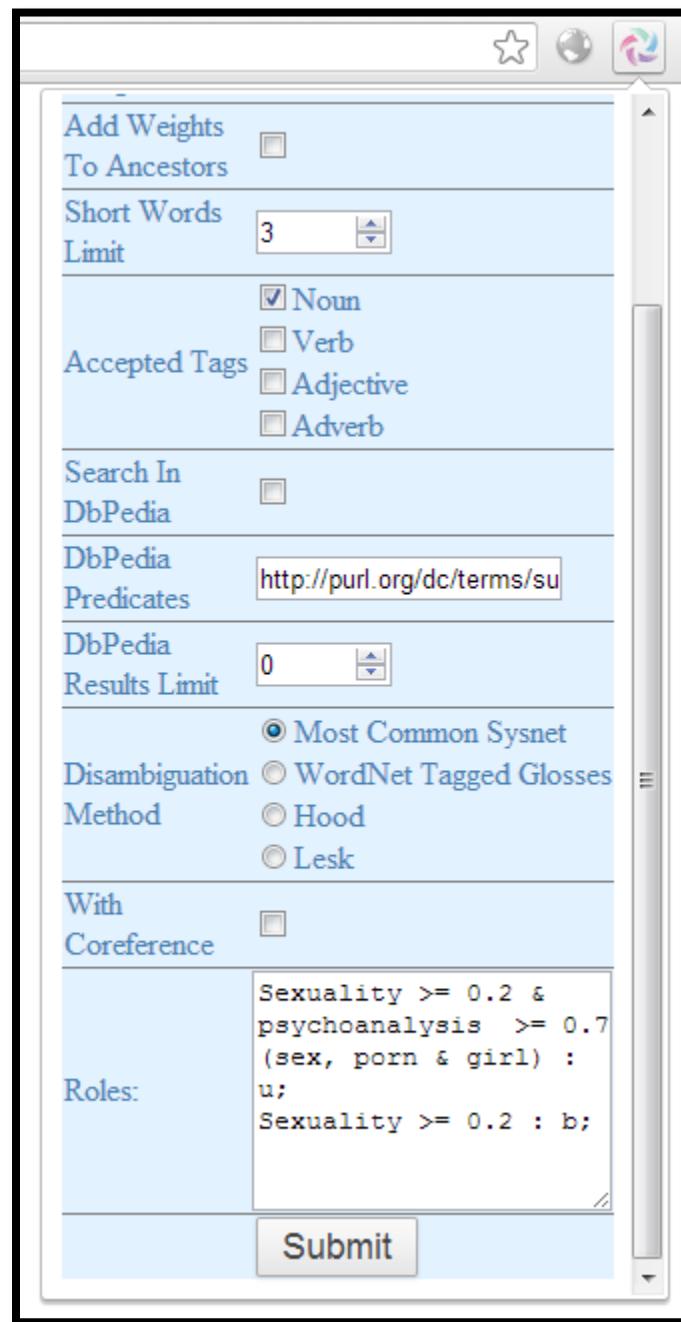
الشكل 66 طلب صفحة ومعالجتها وحجبها

## وضع إعدادات المخدم



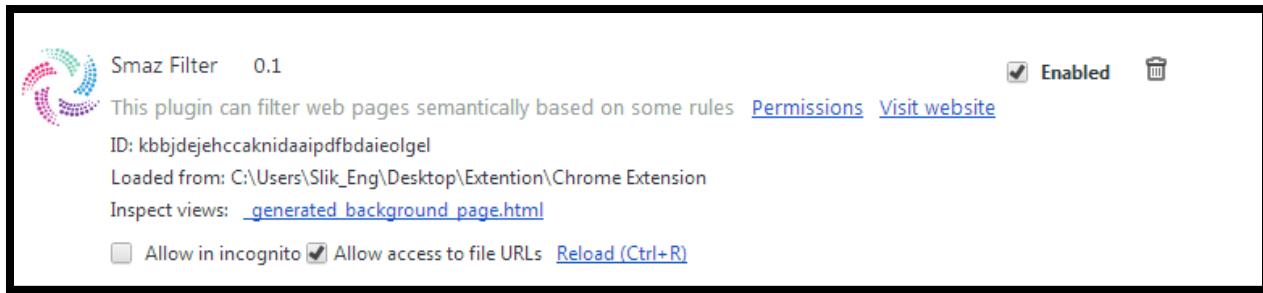
الشكل 67 وضع إعدادات المخدم

### 2.1 واجهة الإضافة التي يتم من خلالها وضع الإعدادات



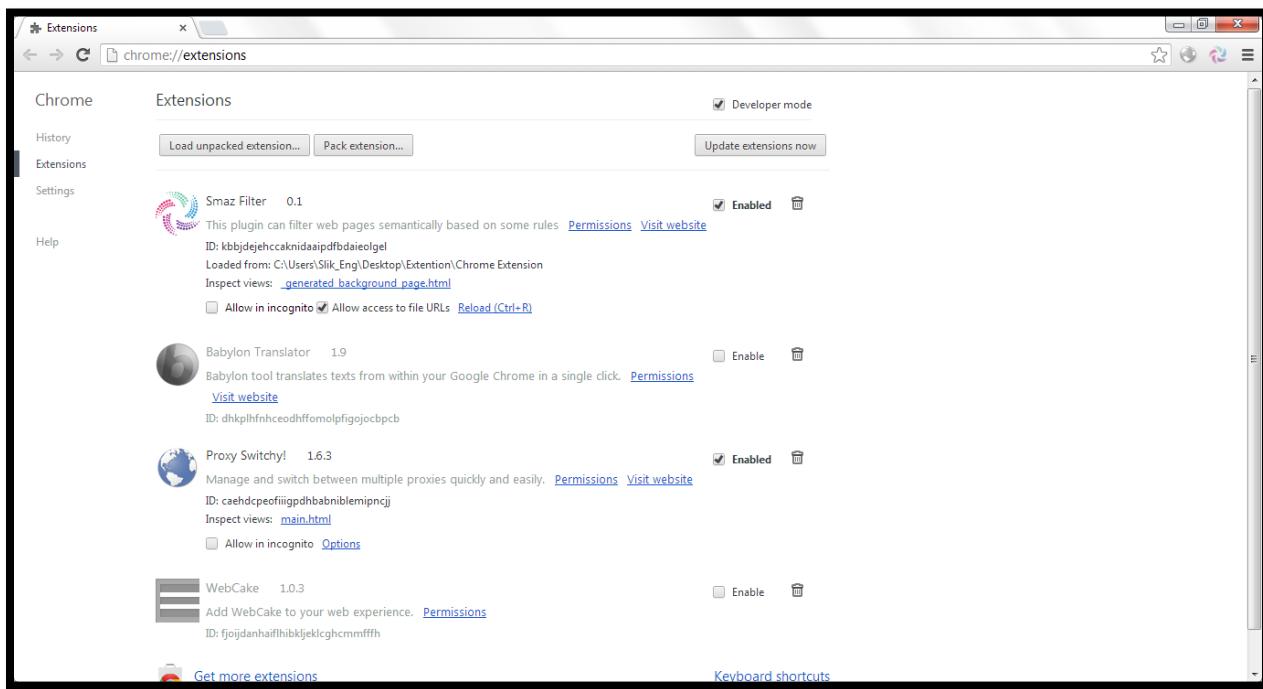
الشكل 48 واجهة الإضافة التي يتم من خلالها وضع الإعدادات

## 2.2 معلومات عن الإضافة



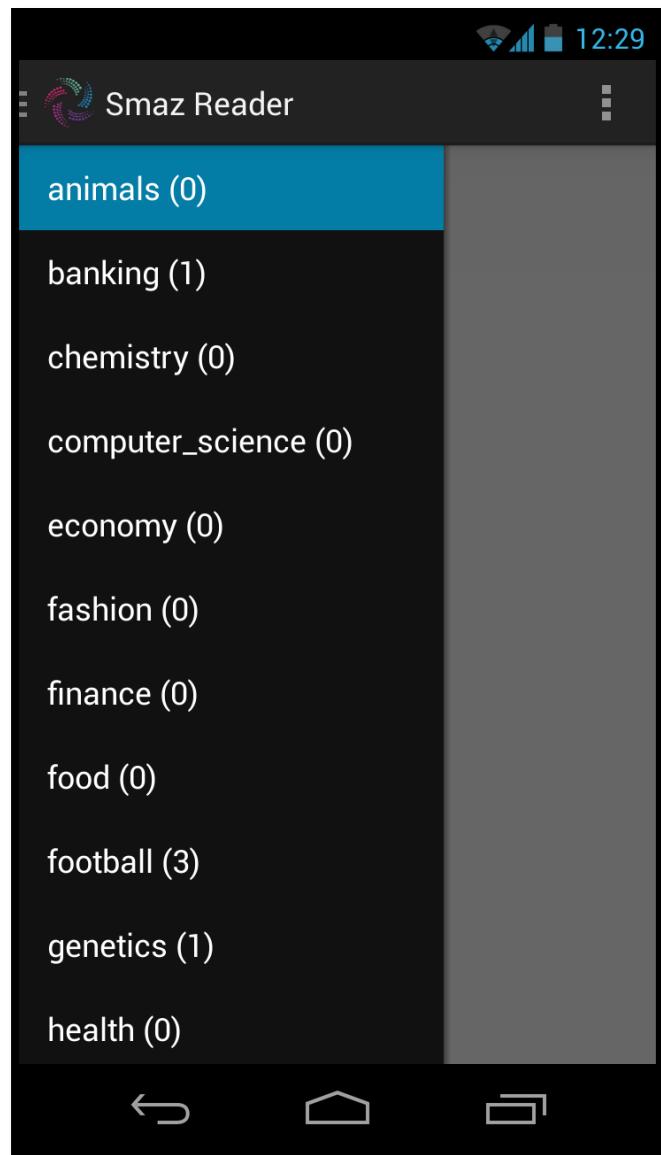
الشكل 69 معلومات عن الإضافة

## 2.3 إضافة Smaz Filter إلى قائمة إضافات المتصفح



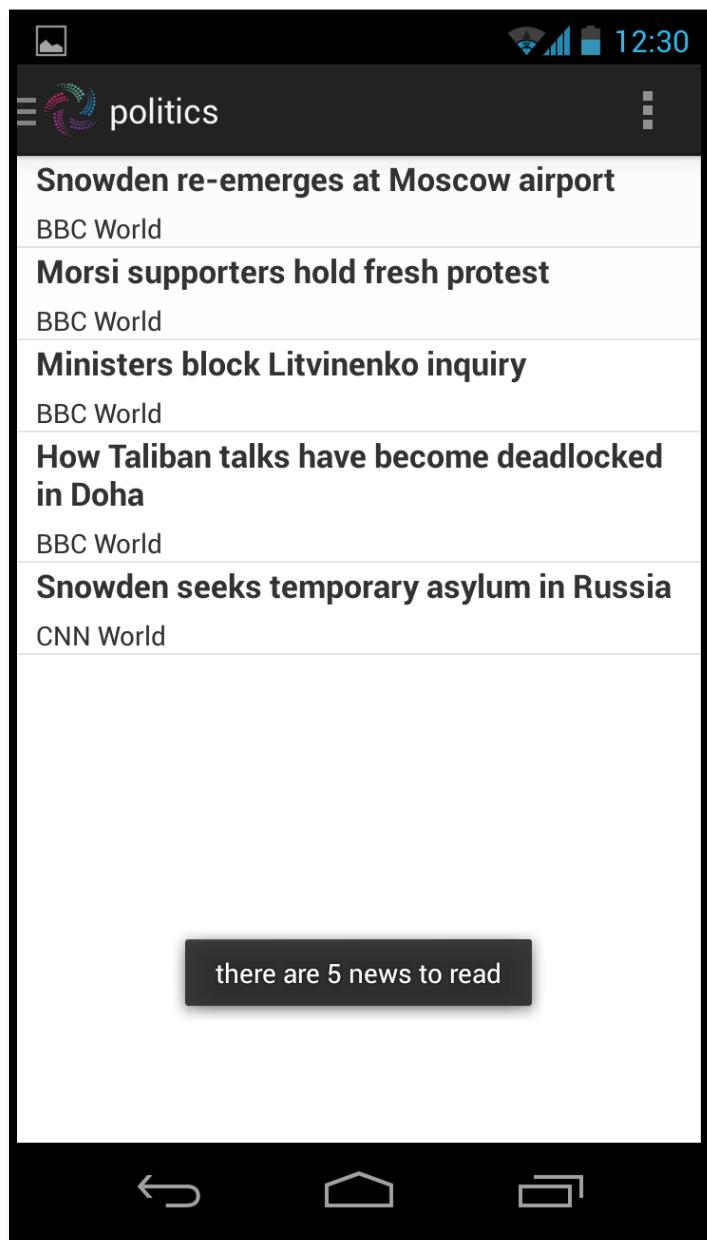
الشكل 70 إضافة Smaz Filter إلى قائمة إضافات المتصفح

### 3.1 واجهة عرض المجالات المختارة مسبقاً



الشكل 71 واجهة عرض المجالات المختارة مسبقاً

### 3.2 واجهة عرض الأخبار الخاصة بـ مجال معين



الشكل 72 واجهة عرض الأخبار الخاصة بـ مجال معين

### 3.3 واجهة إعدادات التطبيق



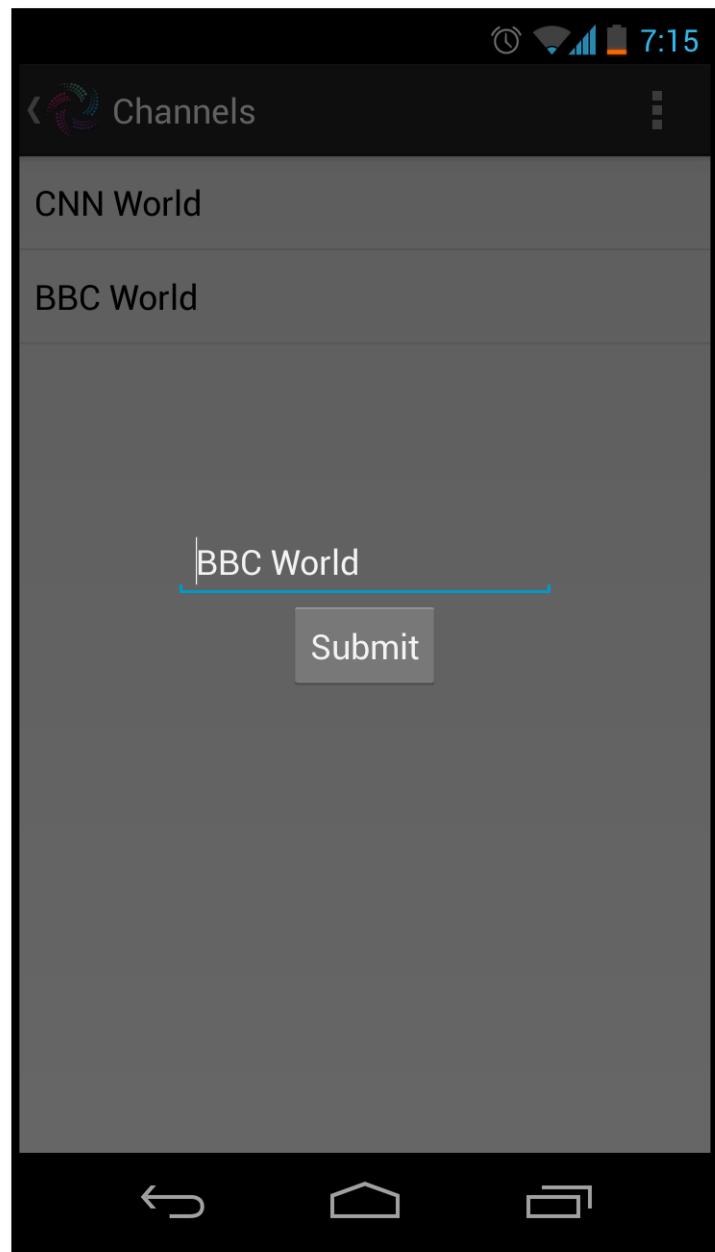
الشكل 73 واجهة إعدادات التطبيق

### 3.4 واجهة عرض قنوات الأخبار



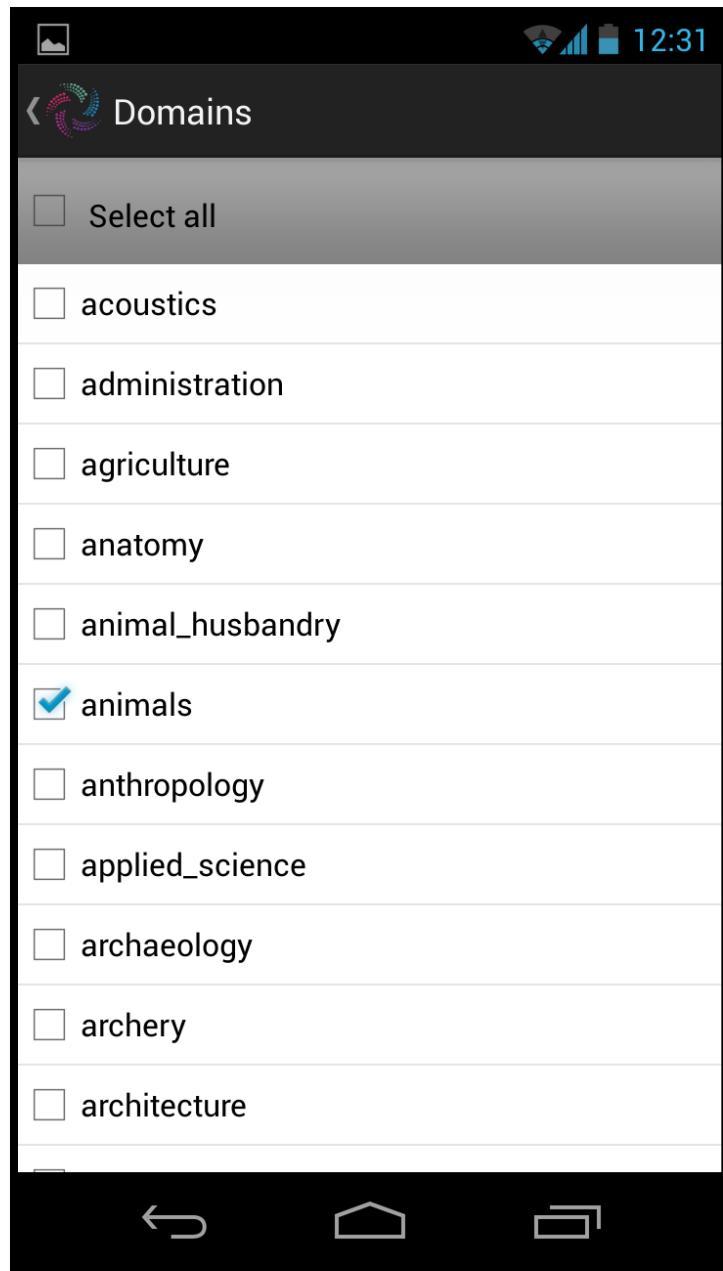
الشكل 74 واجهة عرض قنوات الأخبار

### 3.5 واجهة تعديل قناة أخبار



الشكل 75 واجهة تعديل قناة أخبار جديدة

### 3.6 واجهة توضح اختيار المجالات المطلوب عرضها



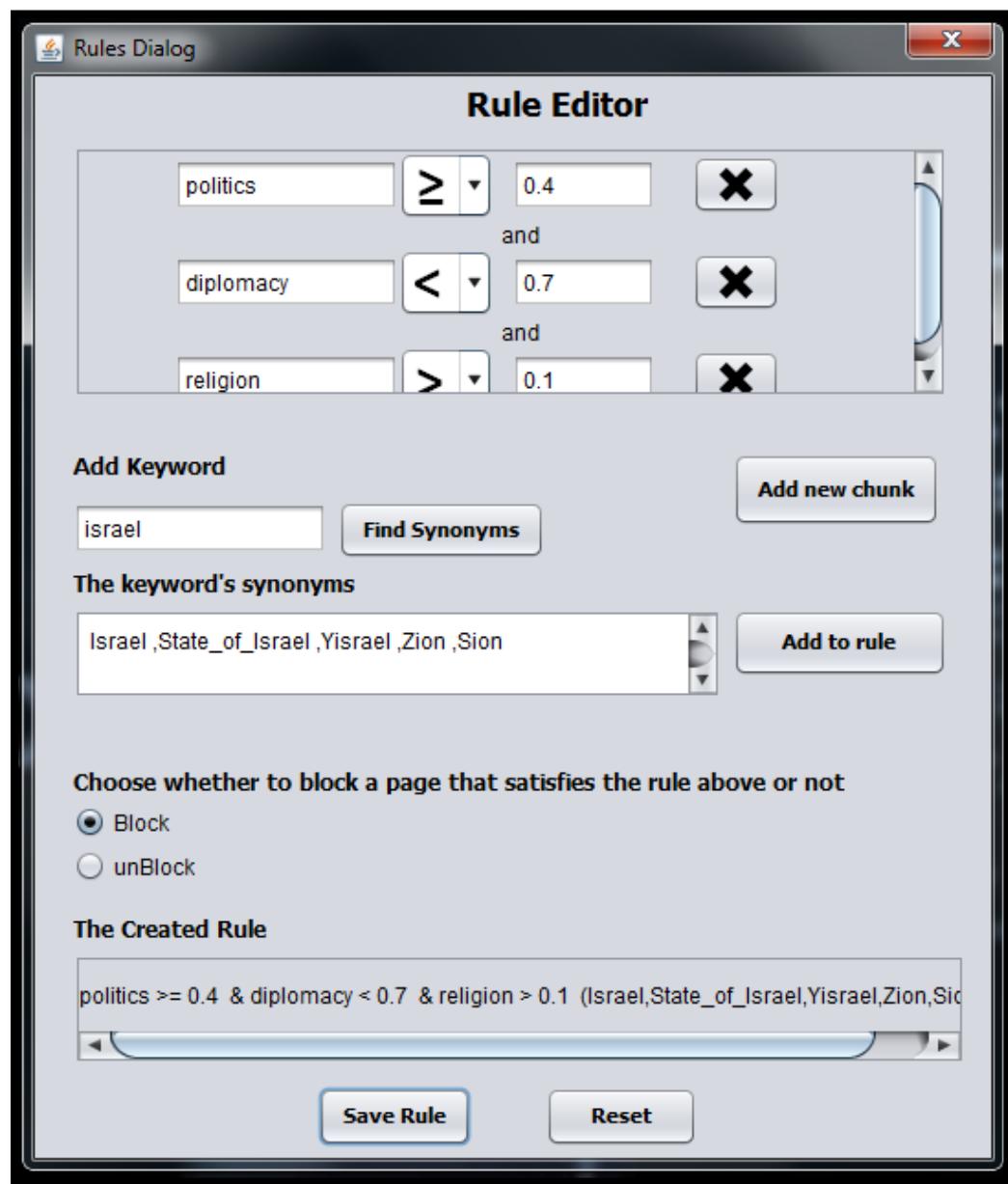
الشكل 76 واجهة توضح اختيار المجالات المطلوب عرضها

### 3.7 واجهة توضع معلومات عن التطبيق



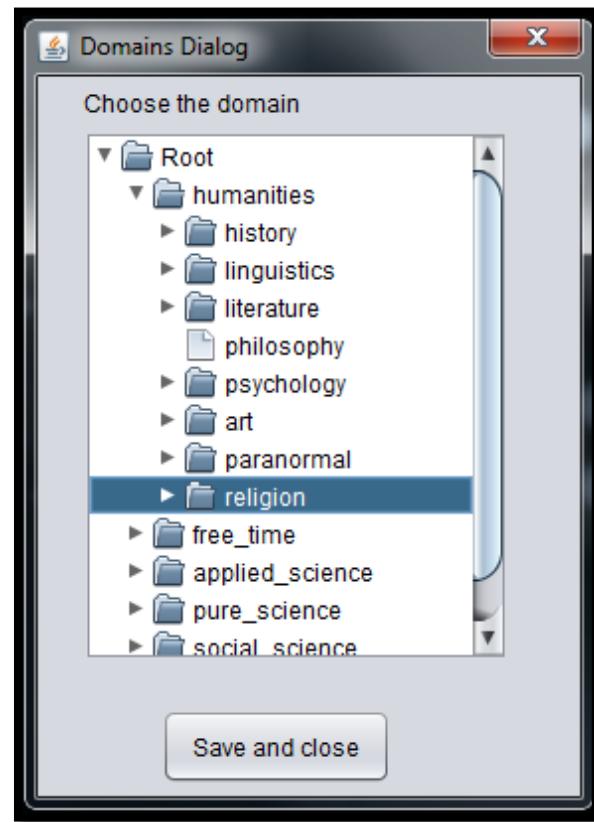
الشكل 77 واجهة توضع معلومات عن التطبيق

#### 4.1 واجهة إضافة قاعدة



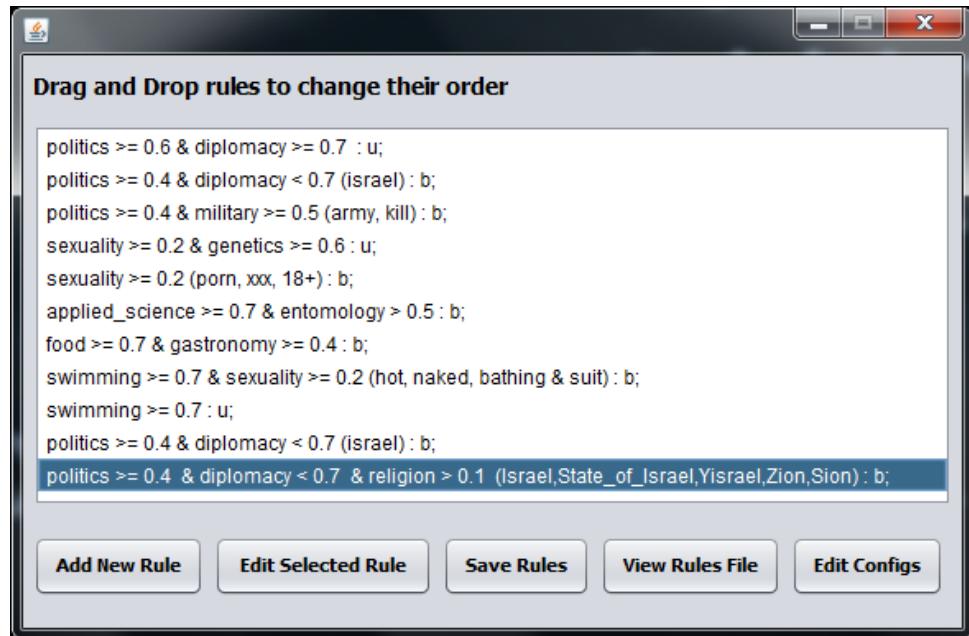
الشكل 78 واجهة إضافة قاعدة

#### 4.2 واجهة اختيار مجال من المجالات وفق الهرمية الموضوعية



الشكل 79 واجهة اختيار مجال من المجالات وفق الهرمية الموضوعية

#### 4.3 واجهة عرض القواعد الموجودة والتعديل عليها



الشكل 80 واجهة عرض القواعد الموجودة والتعديل عليها



## الباب السادس

### مرحلة الامتحانات

---



## 1 مقدمة

يكون التحدي الأكبر في التحقق من صحة عمل النظام بشكل صحيح وتأديته لجميع الوظائف المطلوبة، مما يتضمن إجراء كم كافي من الاختبارات، وبالطبع فإن الطريقة السليمة لاختبار أي نظام يجب أن تتم وفق منهجية صحيحة ومراحل متتالية للتأكد من صحة سير العمل، وبالتالي فقد تم تنفيذ مجموعة من الاختبارات عبر مراحل تطوير النظام ، وخلال مرحلة الاختبارات يجري التأكد من عدم ارتكاب أخطاء خلال مراحل تطوير النظام.

تنص على تتفاوت الأخطاء المكتشفة خلال هذه المرحلة في درجة خطورتها، لأن الخطأ الناتج عن مرحلة التحقيق البرمجي يعتبر من الأخطاء السهلة، بينما الخطأ الناتج عن مرحلة التحليل والتصميم يعتبر من أخطر الأخطاء وذلك لأنه يؤدي للعودة إلى تلك المراحل المبكرة لتصحيحه.

ولابد من التنويه بأن النظام يجب أن يحقق معايير الجودة الداخلية والخارجية.

### 1 الاختبارات عبر مراحل تطوير النظام

#### 1.1 اختبار الوحدة البرمجية

هو الاختبار الذي قام به كل فرد من أفراد المشروع على الوحدة البرمجية التي طورها، ويكون كل فرد بذلك مسؤولاً عن القيام بهذا النوع من الاختبارات.

##### 1.1.1 اختبار التكامل

هو الاختبار الذي يتأكد من واجهات الوحدة البرمجية، وقدرتها على العمل مع بعضها، وقد تم هذا الاختبار عن الانتهاء من أجزاء متصلة من النظام، ولكن قبل أن يتم تجميع النظام ككل.

### **1.1.2 اختبار النظام**

هو الاختبار الذي يتتأكد من أن كل المتطلبات قد تحققت، وجرى هذا الاختبار عندما تم تجميع كل أجزاء النظام، ليتم التحقق من صحة عمل النظام ككل.

### **1.1.3 اختبارات الجودة**

#### **المعايير الخارجية**

##### **1- الصحة**

تعبر عن مدى ملائمة النظام لاحتياجات المستخدم، وقد تم تصميم النظام وفقاً لما اتفق عليه ليلبي الغرض الذي يم لأجله.

##### **2- الوثوقية**

تم مراعات الكثير من الأخطاء التي يمكن أن يرتكبها المستخدم من أجل الحصول على أعلى درجة من الوثائقية.

##### **3- قابلية الصيانة**

تم اعتماد مجموعة من الأمور في النظام لتأمين سهولة الصيانة وإمكانية اكتشاف الخطأ منها:

- لقد تم اعتماد نظام المجرزءات في تصميم النظام مما يسهل اكتشاف مكان الخطأ.
- القيام بعمليات الأرشفة لكل ما يدور في النظام من طلبات مما يسهل إمكانية متابعة العمل.

##### **4- إعادة الاستخدام**

بما أن النظام مكون من مركبات برمجية (مجرزءات)، فإن إعادة استخدام أحد المجرزءات ليس بالأمر الصعب.

## **■ المعابر الداخلية**

### **1- الاتكمال**

تم تحقيق الوظائف المطلوبة من النظام بشكل كامل.

### **2- الاجزائية**

تم فصل أجزاء النظام المختلفة في مركبات مستقلة وظيفياً.



## الباب السابع

### الآفاق المستقبلية، الملحقات والمراجع

---



# الفصل الرابع عشر

## الآفاق المستقبلية

نأمل مستقبلاً أن يتم العمل على تطوير المشروع والاستفادة أكثر من الخدمة التي نقدمها:

- 1 - بناء محرك بحث يجعل نقطة الانطلاق هي المجالات domains، حيث تكون هذا المجالات بمثابة فهرس يتم من خلاله فهرسة صفحات الإنترنت وتوزيعها على المجالات التي تتبع إليها؛
- 2 - دعم اللغة العربية من خلال تطبيق العمل على محتوى باللغة العربية، لم نقوم بهذه العملية لأن هذا العمل يتطلب أنطولوجية باللغة العربية وهذا الشيء غير متوافر حالياً؛
- 3 - الاستعانة بالمزيد من مصادر البيانات والربط مع أنطولوجيات أخرى وبالتالي زيادة الدقة في النتائج فالكلمات التي لا نجدها في wordnet كنا نبحث عنها في dbpedia، نطبع في المستقبل إلى الربط مع أنطولوجيات أكثر.



## الفصل الخامس عشر

### الملحق أ- ملخص المصطلفات

المصطلح باللغة الإنجليزية	اختصار المصطلح	المصطلح باللغة العربية
Children's Internet Protection Act	CIPA	قانون حماية الإنترن特 للطفلة
meta tags		العلامات العامة
proxy server		مخدم بروكسي
chrome extention		إضافة على متصفح كروم
regular expressions		تعابير نظامية
filtering		فلترة
Phrase list		قائمة التعابير
Pornography		صور إباحية
Machine Learning		تعلم الآلة
Dynamic Real Time Rating	DRTR	التقييم الديناميكي في الوقت الحقيقي

add-on		إضافة
Domain		مجال
Grouping		تجميع
Meta Rules		قواعد عامة
Organization		تنظيم
Notification		إشعار
Anti-spam		مضاد لشيء غير مرغوب به
Wordnet		معجم
Wordnet Domains		مجالات المعجم
Synsets		مجموعات
Synonyms		متراادات
Semantic		دلالي
Monosemous		كلمة لها معنى واحد
Polysemous		كلمة لها أكثر من معنى
Gloss		تفسير
Dbpedia		أنطولوجية
Resource Description Framework	RDF	إطار توصيف الموارد
metadata		معلومات عامة وصفية

Extensible Markup Language	<b>XML</b>	لغة التأشير الموسعة
Subject		فاعل
Predicate		فعل
Object		مفعول به
Dataset		بيانات
Simple Protocol and RDF Query Language	<b>SPARQL</b>	لغة الاستعلام على ثلاثيات <b>rdf</b>
Normalization		تطبيع
Coreferencing		رد الضمير إلى أصله
Lemmatization		رد الكلمة إلى أصلها
Disambiguation		إزالة الغموض
Classification		تصنيف
corpus		فضاء
Stopwords		كلمات الغير حاملة للمعنى
Stemming		تجذير
Uniform Resource Identifier	<b>URI</b>	معرف الموارد المنتظم
Uniform Resource Locator	<b>URL</b>	محدد موقع الموارد المنتظم

web service		خدمة ويب
Combination		دمج
Channel		قناة
Remote Method Invocation	RMI	الاستدعاء البعيد للتوابع
Process		معالجة
Request		طلب
Client		زبون
Server		مخدم
Rule		قاعدة
rule editor		محرر القواعد
rule parser		مفسر القواعد
Web Ontology Language	OWL	لغة لبناء الأنطولوجيات
Java Wordnet Interface	JWI	واجهة للتعامل مع معجم wordnet
Splitting		تقسيم النص إلى جمل
Tokenizing		تقسيم الجمل إلى كلمات

جدول 28 مسرد المصطلحات

# الفصل السادس عشر

## الملحق بـ- فهرس المطالعات

---

33.....	الشكل 1 مخطط غانت لمرحلة الدراسة النظرية.....
33.....	الشكل 2 مخطط غانت لمرحلة التحليل.....
34.....	الشكل 3 مخطط غانت لمرحلة التصميم.....
34.....	الشكل 4 مخطط غانت لمرحلة التحقيق .....
35.....	الشكل 5 مخطط غانت لمرحلة الاختبار.....
36.....	الشكل 6 النقاط الأساسية في الدراسة المرجعية.....
38.....	الشكل 7 النقاط الأساسية في مشاريع الفلترة.....
39.....	الشكل 8 النقاط الأساسية في مشاريع الفلترة.....
40.....	الشكل 9 صورة توضح واجهة DansGuardian .....
43.....	الشكل 10 صورة توضح واجهة K9 .....
46.....	الشكل 11 صورة توضح واجهة OpenDNS .....
47.....	الشكل 12 صورة توضح التعامل مع openDNS .....
48.....	الشكل 13 صورة توضح واجهة We-Blocker .....
49.....	الشكل 14 صورة توضح واجهة anti-porn .....
51.....	الشكل 15 صورة توضح واجهة squid .....
54.....	الشكل 16 صورة توضح واجهة safe-squid .....
58.....	الشكل 17 آلية العمل DansGuardian + squid مرحلة 1 .....
58.....	الشكل 18 آلية العمل DansGuardian + squid مرحلة 2 .....

59.....	الشكل 19 آلية العمل DansGuardian + squid مرحلة 3
59.....	الشكل 20 آلية العمل DansGuardian + squid مرحلة 4
60.....	الشكل 21 النقاط الأساسية في فلترة البريد الإلكتروني.....
62.....	الشكل 22 واجهة outlook مرحلة.....
63.....	الشكل 23 واجهة outlook مرحلة 2 .....
65.....	الشكل 24 واجهة outlook مرحلة 3 .....
66.....	الشكل 25 واجهة outlook مرحلة 4 .....
67.....	الشكل 26 واجهة outlook مرحلة 5 .....
68.....	الشكل 27 واجهة outlook مرحلة 6 .....
69.....	الشكل 28 واجهة outlook مرحلة 7 .....
70.....	الشكل 29 واجهة outlook مرحلة 8 .....
74.....	الشكل 30 النقاط الأساسية في مرحلة الدراسة النظرية .....
77.....	الشكل 31 مثال توضيحي عن العلاقات الدلالية في wordnet .....
79.....	الشكل 32 مثال توضيحي عن الارتباطات في wordnet .....
83.....	الشكل 33 مثال عن ثلاثة rdf .....
83.....	الشكل 34 مثال توضيحي عن ثلاثة rdf .....
86.....	الشكل 35 النقاط الأساسية في تحليل وفهم المحتوى.....
89.....	الشكل 36 مثال عن غموض كلمة .....
90.....	الشكل 37 خوارزميات إزالة الغموض.....
94.....	الشكل 38 الدمج بين المجموعات.....
95.....	الشكل 39 أسلوب المقارنة بين المجموعات .....
99.....	الشكل 40 حساب رصيد كل تركيبة.....
101.....	الشكل 41 مثال عن إيجاد جذر hood .....
103.....	الشكل 42 حالات تحديد الجذر في خوارزمية hood .....
112.....	الشكل 43 النقاط الأساسية ضمن مرحلة التحليل .....

الشكل 44 مخطط حالات الاستخدام لخدم البروكسي ..... 119	.....
الشكل 45 مخطط حالات الاستخدام لقارئ الأخبار ..... 120	.....
الشكل 46 إضافة قاعدة إلى قواعد الفلترة ..... 137	.....
الشكل 47 حذف قاعدة ..... 138	.....
الشكل 48 عرض الأخبار التابعة لمجال معين ..... 138	.....
الشكل 49 إضافة قناة أخبار ..... 139	.....
الشكل 50 وضع إعدادات للتطبيق ..... 139	.....
الشكل 51 مخطط الصنوف المفاهيمي لخدم البروكسي ..... 141	.....
الشكل 52 مخطط الصنوف المفاهيمي لقارئ الأخبار ..... 146	.....
الشكل 53 النقاط الأساسية ضمن مرحلة التصميم ..... 148	.....
الشكل 54 مخطط صنوف مخدم البروكسي ..... 150	.....
الشكل 55 مخطط صنوف قارئ الأخبار ..... 157	.....
الشكل 56 مخطط قاعدة المعطيات ..... 167	.....
الشكل 57 النقاط الأساسية ضمن مرحلة التحقيق ..... 168	.....
الشكل 58 النقاط الأساسية في تحقيق النظام ..... 171	.....
الشكل 59 مخطط يوضح آلية عمل خدمة الويب ..... 172	.....
الشكل 60 آلية عمل RMI ..... 173	.....
الشكل 61 مخطط العمل وآلية الاتصال بين المكونات ..... 181	.....
الشكل 62 صفحة الحجب ..... 182	.....
الشكل 63 النقاط الأساسية ضمن واجهات النظام ..... 194	.....
الشكل 64 مخدم icap والبرمجة بداخله ..... 195	.....
الشكل 65 التأكيد من أن المخدم متصل وقدر على معالجة الطلبات ..... 196	.....
الشكل 66 طلب صفحة ومعالجتها وحجبها ..... 197	.....
الشكل 67 وضع إعدادات المخدم ..... 198	.....
الشكل 68 واجهة الإضافة التي يتم من خلالها وضع الإعدادات ..... 199	.....

200.....	الشكل 69 معلومات عن الإضافة .....
200.....	الشكل 70 إضافة إلى قائمة إضافات المتصفح SmaZ Filter .....
201.....	الشكل 71 واجهة عرض المجالات المختارة مسبقا .....
202.....	الشكل 72 واجهة عرض الأخبار الخاصة بمجال معين .....
203.....	الشكل 73 واجهة إعدادات التطبيق .....
204.....	الشكل 74 واجهة عرض قنوات الأخبار.....
205.....	الشكل 75 واجهة تعديل قناة أخبار جديدة .....
206.....	الشكل 76 واجهة توضح اختيار المجالات المطلوب عرضها .....
207.....	الشكل 77 واجهة توضع معلومات عن التطبيق .....
208.....	الشكل 78 واجهة إضافة قاعدة .....
209.....	الشكل 79 واجهة اختيار مجال من المجالات وفق الهرمية الموضوعة .....
210.....	الشكل 80 واجهة عرض القواعد الموجودة والتعديل عليها .....

## الفصل السابع عشر

### الملاحق - فهرس المحتوى

---

80.....	جدول 2 إحصائيات عن معجم wordnet
84.....	جدول 3 إحصائيات عن أنطولوجية DBpedia
107.....	جدول 4 مقارنات بين الخوارزميات الاربعة في إزالة الغموض
121.....	جدول 5 وضع إعدادات النظام
122.....	جدول 6 تعديل إعدادات النظام
123.....	جدول 7 إضافة قاعدة فلترة
125.....	جدول 8 تعديل قاعدة فلترة
126.....	جدول 9 طلب صفحة ويب
127.....	جدول 10 حذف قاعدة فلترة
128.....	جدول 11 عرض محتوى ملف القواعد
129.....	جدول 12 عرض محتوى ملف الإعدادات
130.....	جدول 13 عرض الأخبار التابعة لمجال معين
131.....	جدول 14 تحديد المجالات التي يرغب بعرض الأخبار التابعة لها
132.....	جدول 15 وضع إعدادات التطبيق
133.....	جدول 16 عرض المجالات المتواجدة
134.....	جدول 17 إضافة قناة أخبار

135.....	جدول 18 عرض القنوات المتواجدة .....
135.....	جدول 19 حذف قناة .....
136.....	جدول 20 وضع إعدادات النظام .....
145.....	جدول 21 توصيف الصنوف المفاهيمية لخدم البروكسي .....
147.....	جدول 22 توصيف الصنوف المفاهيمية للقارئ .....
164.....	جدول 23 توصيف الكيان rssChannel .....
165.....	جدول 24 توصيف الكيان Domain .....
165.....	جدول 25 توصيف الكيان RssItem .....
166.....	جدول 26 توصيف الكيان DomainRssItem .....
184.....	جدول 27 اختبار أداء SMAZ Proxy .....
225.....	جدول 28 مسرد المصطلحات .....

## الفصل الثامن عشر

### الملحق س - المراجع

---

- [1] Informations about dansguardian. [Online] <http://dansguardian.org>.
- [2] Informations about K9.[Online]  
[http://en.wikipedia.org/wiki/K9\\_Web\\_Protection](http://en.wikipedia.org/wiki/K9_Web_Protection).
- [3] Informations about opendns. [Online] [www.opendns.com](http://www.opendns.com).
- [4] Informations about Squid. [Online] [en.wikipedia.org/wiki/Squid](http://en.wikipedia.org/wiki/Squid).
- [5] Informations about quintolabs. [Online]  
[http://issues.quintolabs.com/trac/quintolabs\\_qlicap/wiki/BlockSites](http://issues.quintolabs.com/trac/quintolabs_qlicap/wiki/BlockSites).
- [6] Informations about quintolabs. [Online] <http://www.quintolabs.com>.
- [7] Informations about safesquid. [Online] [www.safesquid.com](http://www.safesquid.com).
- [8] Informations about wordnet. [Online] <http://wordnet.princeton.edu>.
- [9] Informations about wordnet domains. [Online] <http://wndomains.fbk.eu>.
- [10] Informations about spotlight - Linking to dbpedia. [Online]  
<http://spotlight.dbpedia.org>.

- [11] Informations about Icap. [Online]  
[http://en.wikipedia.org/wiki/Internet\\_Content\\_Adaptation\\_Protocol.](http://en.wikipedia.org/wiki/Internet_Content_Adaptation_Protocol)
- [12] Informations about features of Icap. [Online] [http://wiki.squid-cache.org/Features/ICAP.](http://wiki.squid-cache.org/Features/ICAP)
- [13] Informations about features of word sense disambiguation. [Online]  
[http://julian.eti.pg.gda.pl/publikacje/wsd.pdf.](http://julian.eti.pg.gda.pl/publikacje/wsd.pdf)
- [14] Informations about features of word sense disambiguation - gloss taged algorithm. [Online] [http://wordnet.princeton.edu/glosstag.html.](http://wordnet.princeton.edu/glosstag.html)
- [15] Informations about features of word sense disambiguation - gloss taged algorithm. [Online] [http://kask.eti.pg.gda.pl/semagloss/index.html.](http://kask.eti.pg.gda.pl/semagloss/index.html)
- [16] Informations about wordnet. [Online]  
[https://ptl.sys.virginia.edu/ptl/members/matthew\\_gerber/software#wordnet.](https://ptl.sys.virginia.edu/ptl/members/matthew_gerber/software#wordnet)
- [17] **Satanjeev Banerjee.** *Adapting the Lesk Algorithm for Word Sense Disambiguation to WordNet*, 2002.
- [18] **Jonas EKEDAHL, koraljka Golub.** *Word sense disambiguation using WordNet and the Lesk algorithm.*
- [19] **Kostas Fragos, Yannis Maistros, Christos Skourlas.** *Word Sense Disambiguation Using WordNet Relations.*