

SeSQL : un moteur de recherche en Python et PostgreSQL

Yohann GABORY — Gaël LE MIGNOT
Pilot Systems

11 juillet 2011

Plan

- 1 Introduction
 - Le besoin initial
 - L'historique du projet
- 2 Fonctionnalités de SeSQL
 - Indexation
 - Recherche
 - Fonctionnalités additionnelles
- 3 Fonctionnement interne
 - Gestion des dépendances
 - Les short queries
 - Quelques optimisations
 - Benchmarks
- 4 Perspectives pour l'avenir
 - Nouvelles fonctionnalités
 - Intégration à d'autres projets
- 5 Conclusion

Introduction

Contexte

Contexte général

- Quotidien Libération
- Utilisation en back-office dans un premier temps
- Utilisation en frontal dans un second temps

Utilisation de la recherche

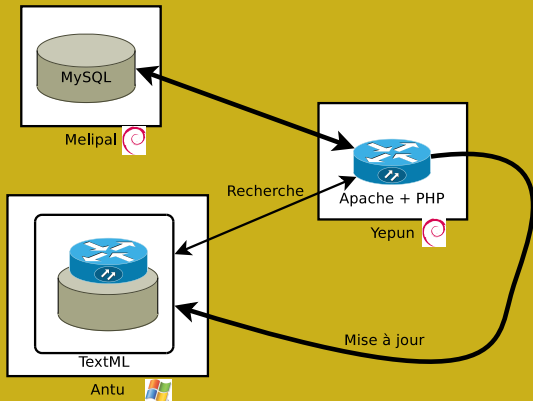
- Navigation
- Recherches simples
- Recherches documentaires

Solution précédente

- Propriétaire, sous Windows (TextML)
- Problèmes de performances et de stabilité

Solution précédente

Schéma d'architecture



Base d'indexation

Volumétrie

- 703 701 articles
- 169 017 pages
- 4 064 478 commentaires (non indexées dans TextML)
- 100k contenus divers

Types d'index

- Recherche en texte plein
- Recherche sur texte exact
- Filtres sur des méta-données : auteurs, catégories, ...
- Tri par date

Interface de recherche

Types de documents

article
 auteur
 blog
 contribution
 diaporama
 émission
 fiche
 page
 media
 nouvelle
 qui-a-dit
 rubrique
 sondage
 tchat

Actions: tous, aucun, article, page

Critères

| | | | | |
|---|--------|----------|------------|------|
| (| Tout | contient | postgresql | OU |
| | Tout | contient | python |) ET |
| | Statut | est | En ligne | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |

Actions: papier, par typologie, par date, par auteur, par rubrique quot., par mot-clés, par num. de page, sur Tout

Tri

Disponibles :

- Pertinence
- Date de création
- Heure de création
- Date de modification
- Heure de modification

>>

Trier par :

- ▼ Date de publication
- ▲ Numéro de page

Actions: réinitialiser, défaut

Affichage

Nombre de résultats:

Afficher les extraits:

Magie:

Actions: défaut

Chercher
Réinitialiser

Version initiale

Contraintes

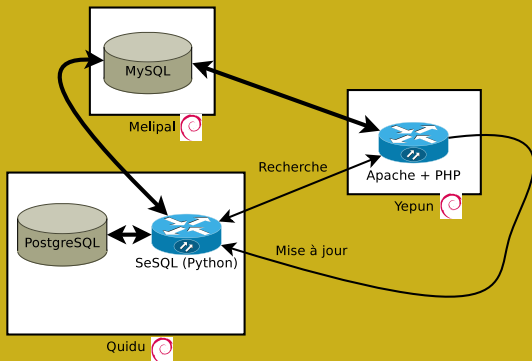
- Devait rester proche de l'architecture existante
- Devait s'interfacer avec du code PHP/MySQL

Solution

- Un webservice en Python
- Une base PostgreSQL séparée
- Une API globalement compatible avec celle de TextML

Première version de SeSQL

Schéma d'architecture



Deuxième version

Contexte

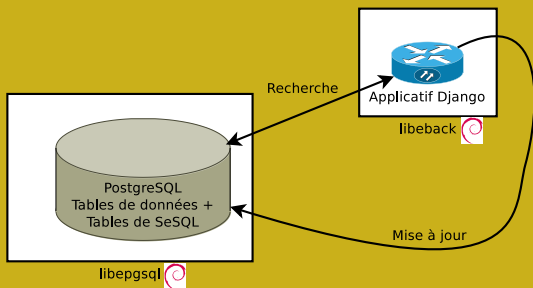
- L'ensemble du site est en cours de migration en Django
- On souhaite s'épargner la lourdeur de l'API XML
- On souhaite rester aussi près que possible de Django

Solution

- Une application Django
- Les recherches s'expriment avec l'objet `Q` de Django
- SeSQL renvoie des objets Django

Deuxième version de SeSQL

Schéma d'architecture



Fonctionnalités de SeSQL

Définition de l'indexation

Types d'index

- Types simples : entiers, dates, ...
- Champs *full text*
- Champ multi-valués (pour des relations par exemple)

Sources des index

- Champs du modèle
- Appel à des méthodes du modèle
- Suivi des relations
- Index composite : concaténation, premier non vide, ...

Intégration avec Django

Configuration de SeSQL

- Une application comme une autre, qui doit être ajoutée dans le `settings.py`
- Nécessite un back-end PostgreSQL
- A son propre fichier de configuration, `sesql_config.py`

Définition des modèles à indexer

- S'effectue via la `TYPE_MAP` dans la configuration
- Par défaut suit les héritages
- Permet de regrouper les contenus cherchés souvent ensembles
- SeSQL utilise un simple signal `post_save`

Un sesql_config minimaliste

```
FIELDS = (ClassField("classname"),
          LongIntField("id"),
          DateTimeField("created_at"),
          FullTextField("user", "user.screen_name"),
          LongIntField("user_id", "user.id"),
          FullTextField("text",
["text"],
primary=True,
dictionary = 'public.lem_french',
          ),
)
```

un exemple d'intégration Rook : Recherche

1 500 000 Tweets en full text

- Des requêtes complexes :

```
tweets = Tweet.objects.filter(  
    user__relation__in = user.relation_set.filter(user_type = "followers")  
).filter(text__icontains = mot).order_by('-created_at')
```

- Simplissime avec SeSQL :

```
ids = request.user.relation_set.filter(user_type="friends").only('id')  
results = longquery(Q(user_id__in = ids) &  
    Q(text__containswords = request.GET['search'] ))
```


Gestion de la lexemisation

Principe

- Prendre le radical des mots
- Peut être très compliqué : cheval, chevaux, chevalet
- Dépend de la langue
- Peut provoquer des effets de bord

Dans SeSQL

- Utilise les *text search configuration* de PostgreSQL
- SeSQL effectue un nettoyage supplémentaire (accents, majuscules, entités HTML, ...)
- Peut être défini de manière différente par index

Description d'une recherche

Utilisation de l'objet Q

- SeSQL utilise l'objet `Q` de Django
- Il permet de définir des recherches complexes, avec des ET, OU, négations, ...
- Chaque élément est composé de : un index, un opérateur et une valeur

Les opérateurs

- Sur du texte : `containswords`, `containsexact`, ...
- Sur des tableaux : `containsall`, `containsany`, ...
- Opérateurs génériques : `plus petit`, `plus grand`, ...

Long query et short query

Short query

- Requêtes pour de la navigation, un portlet, un aperçu, ...
- Limité à un petit nombre (par exemple 50) de résultats
- Ne supporte pas le tri par pertinence
- Extrêmement rapides

Long query

- Supporte la pagination de manière stable
- Envoi le nombre exact de résultats
- Peut être plus lente dans certains cas

Quelques bonus

Le tri

- Sur un index numérique ou de date
- Ou alors par pertinence, avec gestion :
 - de la pondération (le titre compte plus que le corps du texte)
 - de la proximité des mots cherchés
 - de la fréquence des mots cherchés

Les résultats

- `SeSQLResultSet` est un générateur Python, donc lazy
- Renvoi directement les objets Django à l'itération
- Les objets peuvent être de plusieurs types

Et en prime...

Aide au highlight

- Un module spécial est fourni
- Il renvoi la liste des positions, en caractères, des mots ayant été trouvés

SeSQL admin

- Permet d'utiliser SeSQL depuis l'admin Django
- Assez intrusif pour l'instant, donc désactivé par défaut

Historique

- Collecte les historiques de recherche, en option
- Compte le nombre de résultats et la fréquence des recherches

Fonctionnement interne

Gestion des dépendances : le problème

Un exemple

- Dans un module de Quizz, on veut indexer les réponses possibles dans l'objet Quizz
- Facile à faire avec SeSQL et les index composites
- Mais si on modifie une réponse, on ne modifie pas l'objet Quizz

Les conséquences

- On veut pouvoir réindexer les objets liés quand on réindexe un objet
- Mais il peut y en avoir énormément : imaginons qu'on change le nom d'un auteur de 147 789 articles du journal !

Gestion des dépendances : la solution

Le principe

- SeSQL stocke la liste des objets à ré-indexer
- Un *daemon* autonome s'occupe de les ré-indexer, petit à petit

En pratique

- SeSQL ne détecte pas tout seul les objets liés
- Les modèles peuvent implémenter une méthode spécifique, qui indique à SeSQL ce qui doit être réindexé
- Cette méthode peut renvoyer des objets ou des couples (`class, id`).

Les short queries : le problème

Rappel du contexte

- Trouver les 50 articles les plus récentes sur un sujet
- Doit être le plus rapide possible

Les deux *query plans*

- Parcourir l'index sur les dates, et filtrer les articles
- Utiliser l'index sur les articles, puis trier sur les dates

La limite de PostgreSQL

- PostgreSQL a des statistiques sur les mots
- Mais un mot peut être avoir été fréquent mais ne plus l'être

Exemple de plan 1

| Classe | id | Date | Titre | Classe | id | Date | Titre |
|---------|--------|---------------------|---|---------|--------|---------------------|---|
| article | 995730 | 2009-10-07 18:11:18 | referendum en guyane et en martinique sur les institutions le | article | 995727 | 2009-10-07 17:51:06 | inauguration de la diversité à l'ana, avec une promesse clas |
| article | 995729 | 2009-10-07 17:51:15 | ternis: bangla face a gaelguet au deuxième tour a tokyo, inspi | article | 995702 | 2009-10-07 15:21:07 | les roumains dans la rue contre l'avisibilité sur fond de bataille |
| article | 995728 | 2009-10-07 17:51:12 | meridieu d'escrime: quintana s'etre d'affile pour la france et | article | 995680 | 2009-10-07 13:20:37 | les juges francais, enterrime l'annee sur l'affaire de l'arche |
| article | 995727 | 2009-10-07 17:51:08 | inauguration de la diversité à l'ana, avec une promesse clas | | | | |
| article | 995726 | 2009-10-07 17:51:03 | grippe: "petite" epidemie s'etale voire a la bain e | | | | |
| article | 995725 | 2009-10-07 17:42:24 | le port de la burqa, cest davantage une forme de protestation | | | | |
| article | 995724 | 2009-10-07 17:31:06 | kouchner rentra ses declarations contre gads, camara | | | | |
| article | 995723 | 2009-10-07 17:21:11 | paris: le chauffeur de car de tous ne r'ava | | | | |
| article | 995722 | 2009-10-07 17:00:44 | le senat interdit l'usage du salafahna portable dans les ecole | | | | |
| article | 995721 | 2009-10-07 16:51:01 | le cia veut reformer le processus de candidature aux jeux oly | | | | |
| article | 995720 | 2009-10-07 16:50:55 | l'agence francaise antidopage rallume l'incendie autour du to | | | | |
| article | 995719 | 2009-10-07 16:41:24 | athletes michael johnson cons eille a usain bolt de ralentir | | | | |
| article | 995718 | 2009-10-07 16:41:17 | l'avis: une affiche anti-minarets de la droite dure "amse a la h | | | | |
| article | 995717 | 2009-10-07 16:32:20 | le senat interdit l'usage du portable a l' ecole | | | | |
| article | 995716 | 2009-10-07 16:32:15 | corruption en espagne: la droite arlabousses, rajoy, eige d'e | | | | |
| article | 995715 | 2009-10-07 16:21:51 | us: salana maintenu en poste un peu plus longtemps que pir | | | | |
| article | 995714 | 2009-10-07 16:21:40 | restauration: la base de la naa permettant de crer 0 000 en | | | | |
| article | 995713 | 2009-10-07 16:21:36 | les abs equipe de la legende s'entend et l'uae celebre sa mi | | | | |
| article | 995712 | 2009-10-07 16:11:16 | promesse vis de du roi abdallah d'arabie s'annonce en sau | | | | |
| article | 995711 | 2009-10-07 16:01:06 | clausstream confronte les gouverns bataillon parole contre | | | | |
| article | 995710 | 2009-10-07 16:01:01 | le ministre de l'agriculture, huno le maire , ampacha d'enter | | | | |
| article | 995709 | 2009-10-07 16:00:51 | gouvernement et opposition ouvent la guerre des jeux en lig | | | | |
| article | 995708 | 2009-10-07 15:51:05 | vacue liam , s'ous pression pour ign er le triste de la banne | | | | |
| article | 995707 | 2009-10-07 15:41:15 | decouvre du "sageur des anneaux " du systeme relaine au | | | | |
| article | 995706 | 2009-10-07 15:38:02 | les albatros trahis par la video surveillance | | | | |
| article | 995705 | 2009-10-07 15:34:49 | le cas se-ete afghan d'abama | | | | |
| article | 995704 | 2009-10-07 15:32:08 | formule 1: le polonais robert lubica che renait en 2010 | | | | |
| article | 995703 | 2009-10-07 15:21:12 | avis: le syndrome du stanc , bulgi plait pour l' homme er | | | | |
| article | 995702 | 2009-10-07 15:21:07 | les troups des us cont re l'austrite sur fond de bataill | | | | |
| article | 995701 | 2009-10-07 15:20:59 | donald tusk annonce un remaniement gouvernemental en pol | | | | |
| article | 995700 | 2009-10-07 15:11:20 | canons a eau , gaz lacrimogene contre des manifes tants anti | | | | |
| article | 995699 | 2009-10-07 15:01:59 | meridieu d'escrime: les seakites francais , en finale contre la fr | | | | |
| article | 995698 | 2009-10-07 15:01:58 | corse du nerd : l'anneux spacial francais , je et lang , annonce r | | | | |
| article | 995697 | 2009-10-07 14:54:32 | abdallah , ly a une recrudescence des assauts de grandes | | | | |
| article | 995696 | 2009-10-07 14:54:04 | miterand vis e pours en au str , sur le touris me sexuel | | | | |
| article | 995695 | 2009-10-07 14:41:36 | madagascar : incertitudes sur le consensus de sorte de cris e | | | | |
| article | 995694 | 2009-10-07 14:41:33 | le navire amrai francais , attaque par des pirates somalien s | | | | |
| article | 995693 | 2009-10-07 14:21:11 | lang , le president ahmadinejad qualifie les nagociations , de gar | | | | |
| article | 995692 | 2009-10-07 14:21:09 | le ministre de la culture francais miterand campagne par une pr | | | | |
| article | 995691 | 2009-10-07 14:10:53 | khmen rouges : le tribunal che des ministres cambodgien s co | | | | |
| article | 995690 | 2009-10-07 14:04:04 | le laurat art a restitu er des statue s a l'egypte | | | | |
| article | 995689 | 2009-10-07 14:01:08 | grice : le premier ministre papanandreu s' engage a modern e | | | | |
| article | 995688 | 2009-10-07 14:01:06 | grand paris : le texte du conseil des ministres meccaniste de g | | | | |
| article | 995687 | 2009-10-07 13:50:49 | graves malnutritions dans une maison de retrete a bagnolet | | | | |
| article | 995686 | 2009-10-07 13:50:44 | un camp de migrants d'antenne a cais , besson confirme du | | | | |
| article | 995685 | 2009-10-07 13:41:05 | dopage tour de france 2008: les nouvelles analyses l'annee | | | | |
| article | 995684 | 2009-10-07 13:41:02 | afghans lan , un militaire espagnol tu , cing autres blesses d | | | | |
| article | 995683 | 2009-10-07 13:30:52 | syrie : premier vis de du roi abdallah d'arabie s' annonce , apre | | | | |
| article | 995682 | 2009-10-07 13:20:43 | une majorite de francais s' declare favorable aux ports me | | | | |
| article | 995681 | 2009-10-07 13:20:39 | lang , le chef de la police d ement tour violence : dans une pr | | | | |
| article | 995680 | 2009-10-07 13:20:37 | les juges francais , ent terme l'annee sur l'affaire de l'arche | | | | |
| article | 995679 | 2009-10-07 13:11:17 | l'egypte se pend s'a capitulation avec le laurat , pour recup | | | | |
| article | 995678 | 2009-10-07 12:50:54 | idiane avec le "vaincu": le 10 novembre a amens , au profit | | | | |
| article | 995677 | 2009-10-07 12:46:48 | deficit budgetaire , rappele neuf nouveaux pays a l' ordre | | | | |
| article | 995676 | 2009-10-07 12:45:18 | fin de cris a madagascar | | | | |
| article | 995675 | 2009-10-07 12:41:37 | cristiano ronaldo au casu d'une "guerre entre so rcien " | | | | |
| article | 995674 | 2009-10-07 12:32:23 | france frappe s'op pos les bleus visent le score floue | | | | |

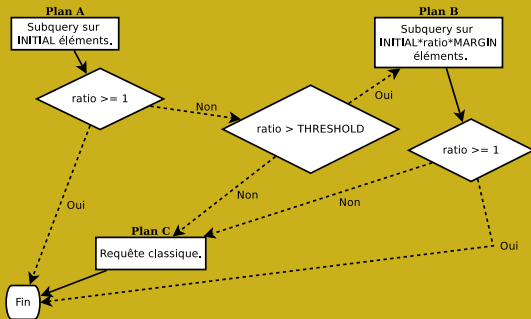
Exemple de plan 2

| Classe | id | Date | Titre |
|---------|--------|---------------------|--|
| article | 959730 | 2009-10-07 18:11:18 | (referendum en guyane et en martinique sur les institutions le |
| article | 959729 | 2009-10-07 17:45:45 | senegal : bangka a gaouquet au deuxième tour a takou, mont |
| article | 959728 | 2009-10-07 17:51:12 | mondoulu d'es crime : ouattara s'écrit d'affaire pour la france et |
| article | 959727 | 2009-10-07 17:51:06 | inauguration de la diouana à fatick, avec une premiere classe |
| article | 959726 | 2009-10-07 17:51:03 | grappe "perle" episteme : table voire a la baïse |
| article | 959725 | 2009-10-07 17:42:24 | le port de la burqa, cast devant une forme de protes tator |
| article | 959724 | 2009-10-07 17:39:00 | kouchner reitere ses declarations contre dadi, camara |
| article | 959723 | 2009-10-07 17:31:11 | paris : le chauffeur de car de toubon me rait plus de france et |
| article | 959722 | 2009-10-07 17:21:46 | le senat interdit l'usage du talaphana portable dans les aéro |
| article | 959721 | 2009-10-07 16:51:01 | le dia-vert reforme le processus de candidature aux presy |
| article | 959720 | 2009-10-07 16:50:55 | l'agence française antiterrorisme refuse l'incendie autour du t |
| article | 959719 | 2009-10-07 16:41:24 | adlets me michael ohgon conseille a usain bolt de valent |
| article | 959718 | 2009-10-07 16:38:17 | l'usage d'une affiche antimilitariste de la droite dure "ants a la |
| article | 959717 | 2009-10-07 16:32:20 | le senat interdit l'usage du portable à l'aéro |
| article | 959716 | 2009-10-07 16:32:15 | corruption en espagne : la droite à l'abusives, rajoy, elige d |
| article | 959715 | 2009-10-07 16:21:51 | us : salana maintenu en poste un peu plus longtemps que gar |
| article | 959714 | 2009-10-07 16:21:40 | restauration : la baïse de la tua permettrait de créer 0 000 em |
| article | 959713 | 2009-10-07 16:21:35 | les ab : que sera le jupron et sera célèbre ce a mité |
| article | 959712 | 2009-10-07 16:11:30 | promesse verte du roi abdallah d'arrêter l'insulte en syrie |
| article | 959711 | 2009-10-07 16:01:06 | classement confronte : les preneurs battent parole contre |
| article | 959710 | 2009-10-07 16:01:01 | le ministre de l'agriculture, brune le maire, amecha d'entre |
| article | 959709 | 2009-10-07 16:00:51 | gouvernement et opposition ouvrent la guerre des jeus en lig |
| article | 959708 | 2009-10-07 15:51:05 | vaccins h1n1 : sous pression pour signer le traite de la banne |
| article | 959707 | 2009-10-07 15:41:15 | décapitée du "saigneur des anneaux" du système bancaire au |
| article | 959706 | 2009-10-07 15:38:02 | les albatros trahis par la vidéosurveillance |
| article | 959705 | 2009-10-07 15:34:49 | la case-rite afghan d'gabana |
| article | 959704 | 2009-10-07 15:32:08 | formulaire 1 : le polonais robert lukbica chez ranauh en 2010 |
| article | 959703 | 2009-10-07 15:21:42 | avec : le syndrome du tbanic, bulle piteuse pour l'homme et la |
| article | 959702 | 2009-10-07 15:21:02 | les sommes d'ant, le nez contre l'austrite sur fond de batall |
| article | 959701 | 2009-10-07 15:20:59 | donald tusk annonce un remaniement gouverneme ntal en pol |
| article | 959700 | 2009-10-07 15:11:20 | canons à eau, gaz lacrimogène contre des manifestants anti |
| article | 959699 | 2009-10-07 15:01:59 | mondoulu d'es crime : les epistes français en finale contre la |
| article | 959698 | 2009-10-07 15:01:58 | corée du nord : l'envoye special français, jack lang, annonce s |
| article | 959697 | 2009-10-07 14:54:32 | shahing : fly a une recrudescence des operations de grande |
| article | 959696 | 2009-10-07 14:54:04 | miterand vivé pours en azul, sur le touris me sexuel |
| article | 959695 | 2009-10-07 14:41:30 | madagascar : incertitudes sur le consensus de sorte de cite |
| article | 959694 | 2009-10-07 14:41:33 | la navire amiral français, attaque par des pirates somaliens |
| article | 959693 | 2009-10-07 14:21:11 | kau : le president ahmadadjan qualifie les negociations, de gar |
| article | 959692 | 2009-10-07 14:21:09 | le ministre de la culture frederic miterrand ramapo pui avec |
| article | 959691 | 2009-10-07 14:10:53 | thmes rouges : le tribunal cite des ministres cambodgiens con |
| article | 959690 | 2009-10-07 14:04:04 | le louve prêt a restituer des viels a l'agette |
| article | 959689 | 2009-10-07 14:01:08 | grèce : le premier ministre papandreaus s'engage a modernis |
| article | 959688 | 2009-10-07 14:01:06 | grand paris : le texte du conseil des ministres meconforte de |
| article | 959687 | 2009-10-07 13:50:49 | grave minuitisme : dans une maison du terville a bayonne |
| article | 959686 | 2009-10-07 13:50:44 | un camp de migrants demantelle a calais, besson confirme de |
| article | 959685 | 2009-10-07 13:41:05 | depaget-tour de france 2008 : les nouvelles analyses restait |
| article | 959684 | 2009-10-07 13:41:02 | afghanistan : un militaire espagnol tue, cinq autres blesses de |
| article | 959683 | 2009-10-07 13:30:52 | syrie : premiere visite du roi abdallah d'arabia s soulève apres |
| article | 959682 | 2009-10-07 13:20:43 | une manibre de français se declare favorable aux ports mico |
| article | 959681 | 2009-10-07 13:20:39 | kau : le chef de la police demont tout viol dans une prison de |
| article | 959680 | 2009-10-07 13:20:37 | les juges français, ont termine l'enquete sur l'affaire de farche |
| article | 959679 | 2009-10-07 13:11:17 | le grece suspend sa cooperation avec le louve pour recuati |
| article | 959678 | 2009-10-07 12:50:54 | ridane avec le "vaincu" : le 10 novembre a amens, au profit |
| article | 959677 | 2009-10-07 12:46:10 | deficit budgétaire : rappelle neuf nouveaux pays à l'ordre |
| article | 959676 | 2009-10-07 12:45:18 | fin de crise a madagascar |
| article | 959675 | 2009-10-07 12:41:37 | cihsano ranala au casus d'une "guerre entre sorcier" |
| article | 959674 | 2009-10-07 12:32:23 | france-fesse a amens : les houp vivant le score l'houe |

| Classe | id | Date | Titre |
|---------|--------|---------------------|--|
| article | 959729 | 2009-10-07 17:51:19 | senegal : bangka a gaouquet au deuxième tour a takou, mont |
| article | 959728 | 2009-10-07 17:51:12 | mondoulu d'es crime : ouattara s'écrit d'affaire pour la france et |
| article | 959727 | 2009-10-07 17:51:06 | inauguration de la diouana à fatick, avec une premiere classe |
| article | 959726 | 2009-10-07 17:51:03 | grappe "perle" episteme : table voire a la baïse |
| article | 959725 | 2009-10-07 17:42:24 | le port de la burqa, cast devant une forme de protes tator |
| article | 959724 | 2009-10-07 17:39:00 | kouchner reitere ses declarations contre dadi, camara |
| article | 959723 | 2009-10-07 17:31:11 | paris : le chauffeur de car de toubon me rait plus de france et |
| article | 959722 | 2009-10-07 17:21:46 | le senat interdit l'usage du talaphana portable dans les aéro |
| article | 959721 | 2009-10-07 16:51:01 | le dia-vert reforme le processus de candidature aux presy |
| article | 959720 | 2009-10-07 16:50:55 | l'agence française antiterrorisme refuse l'incendie autour du t |
| article | 959719 | 2009-10-07 16:41:24 | adlets me michael ohgon conseille a usain bolt de valent |
| article | 959718 | 2009-10-07 16:38:17 | l'usage d'une affiche antimilitariste de la droite dure "ants a la |
| article | 959717 | 2009-10-07 16:32:20 | le senat interdit l'usage du portable à l'aéro |
| article | 959716 | 2009-10-07 16:32:15 | corruption en espagne : la droite à l'abusives, rajoy, elige d |
| article | 959715 | 2009-10-07 16:21:51 | us : salana maintenu en poste un peu plus longtemps que gar |
| article | 959714 | 2009-10-07 16:21:40 | restauration : la baïse de la tua permettrait de créer 0 000 em |
| article | 959713 | 2009-10-07 16:21:35 | les ab : que sera le jupron et sera célèbre ce a mité |
| article | 959712 | 2009-10-07 16:11:30 | promesse verte du roi abdallah d'arrêter l'insulte en syrie |
| article | 959711 | 2009-10-07 16:01:06 | classement confronte : les preneurs battent parole contre |
| article | 959710 | 2009-10-07 16:01:01 | le ministre de l'agriculture, brune le maire, amecha d'entre |
| article | 959709 | 2009-10-07 16:00:51 | gouvernement et opposition ouvrent la guerre des jeus en lig |
| article | 959708 | 2009-10-07 15:51:05 | vaccins h1n1 : sous pression pour signer le traite de la banne |
| article | 959707 | 2009-10-07 15:41:15 | décapitée du "saigneur des anneaux" du système bancaire au |
| article | 959706 | 2009-10-07 15:38:02 | les albatros trahis par la vidéosurveillance |
| article | 959705 | 2009-10-07 15:34:49 | la case-rite afghan d'gabana |
| article | 959704 | 2009-10-07 15:32:08 | formulaire 1 : le polonais robert lukbica chez ranauh en 2010 |
| article | 959703 | 2009-10-07 15:21:42 | avec : le syndrome du tbanic, bulle piteuse pour l'homme et la |
| article | 959702 | 2009-10-07 15:21:02 | les sommes d'ant, le nez contre l'austrite sur fond de batall |
| article | 959698 | 2009-10-07 15:01:58 | corée du nord : l'envoye special français, jack lang, annonce s |
| article | 959697 | 2009-10-07 14:54:32 | shahing : fly a une recrudescence des operations de grande |
| article | 959695 | 2009-10-07 14:41:36 | madagascar : incertitudes sur le consensus de sorte de cite |
| article | 959691 | 2009-10-07 14:10:53 | thmes rouges : le tribunal cite des ministres cambodgiens con |
| article | 959690 | 2009-10-07 14:04:04 | le louve prêt a restituer des viels a l'agette |
| article | 959687 | 2009-10-07 13:50:49 | grave minuitisme : dans une maison du terville a bayonne |
| article | 959685 | 2009-10-07 13:41:05 | depaget-tour de france 2008 : les nouvelles analyses restait |
| article | 959684 | 2009-10-07 13:41:02 | afghanistan : un militaire espagnol tue, cinq autres blesses de |
| article | 959679 | 2009-10-07 13:11:17 | l'agette suspend sa cooperation avec le louve pour recuati |
| article | 959676 | 2009-10-07 12:45:18 | fin de crise a madagascar |
| article | 959675 | 2009-10-07 12:41:37 | cihsano ranala au casus d'une "guerre entre sorcier" |

L'heuristique

L'algorithme



Recherche par texte exact

Problème

- PostgreSQL n'a pas d'index supportant ça
- Faire un `LIKE` sur toute la base est bien trop lent

La solution

- On commence pas filtrer, via l'index, en texte approché
- Puis on refiltre avec un `LIKE` sur ce qui matché

La cas *France 2*

- Beaucoup d'article contiennent les deux mots
- Mais peu le texte exact

Les partitions

Problème

- On veut indexer du contenu très massif (les commentaires)
- On ne veut pas impacter les performances du reste

La solution

- PostgreSQL supporte l'héritage de tables
- On indexe les commentaires dans une table à part
- On ne cherche que dans une sous-table, si possible
- Si non, on cherche dans la table maîtresse

There are three kinds of lies ...

| | <u>TextML</u> | <u>SeSQL</u> | <u>SeSQL</u> | <u>SeSQL</u> | <u>SeSQL</u> | <u>SeSQL</u> | |
|-------------------------------|---------------|--------------|--------------|-------------------|--------------|--------------|-------|
| Partitions | Non | Non | Non | <u>PostgreSQL</u> | <u>SeSQL</u> | <u>SeSQL</u> | |
| Contributions | Non | Non | Oui | Oui | Oui | Oui | |
| Mémoire | 4go | 8go | 12go | 12go | 12go | 4go | |
| Nombre de requêtes | 19280 | 19280 | 19280 | 19280 | 19280 | 19280 | |
| Temps moyen | 0,955 | 0,031 | 0,090 | 0,065 | 0,018 | 0,033 | |
| >20s | 0 | 0 | 0 | 0 | 0 | 0 | |
| >10s | 14 | 0 | 0 | 1 | 0 | 1 | |
| >5s | 154 | 5 | 2 | 7 | 0 | 8 | |
| >2s | 1002 | 10 | 41 | 7 | 0 | 33 | |
| >1s | 7221 | 31 | 484 | 9 | 0 | 69 | |
| 5 plus lentes requêtes | | | | | | | |
| | 1 | 14,46 | 8,69 | 9,08 | 10,6 | 0,82 | 10,23 |
| | 2 | 14,2 | 7,93 | 7,46 | 5,49 | 0,59 | 7,96 |
| | 3 | 13,75 | 5,57 | 4,27 | 5,35 | 0,55 | 7,2 |
| | 4 | 13,61 | 5,16 | 3,3 | 5,3 | 0,54 | 6,49 |
| | 5 | 13,12 | 5,15 | 3,25 | 5,29 | 0,53 | 6,39 |

... lies, damn lies and benchmarks

| | <u>TextML</u> | <u>Sesql</u> | <u>Sesql</u> | <u>Sesql</u> | <u>Sesql</u> | <u>Sesql</u> | <u>Sesql</u> |
|-------------------------------|---------------|--------------|--------------|--------------|--------------|--------------|--------------|
| Nombre de recherches | 19280 | 19280 | 19280 | 19280 | 19280 | 19280 | 19280 |
| Recherches simultanées | n/a | 1 | 2 | 4 | 1 | 2 | 4 |
| Nombre d'insertions | n/a | 0 | 0 | 0 | 19306 | 12157 | 15657 |
| Insertions simultanées | n/a | 0 | 0 | 0 | 1 | 1 | 3 |
| Temps total | n/a | 329,49 | 259,47 | 244,88 | 1594,88 | 1345,59 | 1519,24 |
| Temps moyen | 0,955 | 0,017 | 0,022 | 0,041 | 0,049 | 0,074 | 0,171 |
| >10s | 14 | 0 | 0 | 0 | 0 | 2 | 4 |
| >5s | 154 | 0 | 0 | 0 | 9 | 11 | 56 |
| >2s | 1002 | 0 | 0 | 0 | 70 | 104 | 310 |
| >1s | 7221 | 0 | 0 | 8 | 168 | 235 | 643 |
| 5 plus lentes requêtes | | | | | | | |
| 1 | 14,46 | 0,61 | 0,73 | 1,51 | 6,26 | 17,4 | 17,27 |
| 2 | 14,2 | 0,54 | 0,71 | 1,41 | 6,21 | 11,81 | 13,37 |
| 3 | 13,75 | 0,48 | 0,64 | 1,38 | 6,18 | 9,87 | 11,47 |
| 4 | 13,61 | 0,46 | 0,63 | 1,18 | 5,81 | 8,17 | 10,28 |
| 5 | 13,12 | 0,44 | 0,62 | 1,12 | 5,77 | 7,27 | 8,38 |

Perspectives pour l'avenir

Quelques idées de nouvelles fonctionnalités

Suggestions de recherche

- À partir de l'historique
- Propose des recherches proches, fréquentes et ayant donné de nombreux résultats

Gestion améliorée du multilingue

- Ajout d'une détection automatique de la langue
- Choix de la bonne lexemisation en fonction de la langue
- Pouvoir restreindre la recherche par langue

Possibilité d'intégration à d'autres projets

Hors Django

- Implémenter une API pour avoir un mode webservice
- Implémenter une API compatible avec le ZCatalog pour Zope

Avec d'autres projets

- Intégration possible avec haystack ?
- Avec des moteurs de classification ?

Conclusion

Situation actuelle


- SeSQL disponible en GPL sur bitbucket
- Utilisé en production sur Libération (front et back)
- Utilisé sur d'autres projets comme Rook

En attendant

Remerciements

- aux communautés PostgreSQL, Python et Django
- à Libération de nous avoir fait confiance
- à Jérôme Petazzoni qui a contribué à la conception

La page de pub

- Pilot Systems, société de services en logiciels libres :
<http://www.pilotsystems.net>
- Slides en licence CC-BY-Sa 
- <http://contributions.pilotsystems.net/>

Des questions ?