ATLAS : Outil d'automatisation du maillage interne d'un site web basé sur les techniques d'embeddings

Cyril Wolfangel*, Steve Bellart*
Arnaud Deleruyelle*

*Talan, 14 Rue Pergolèse, 75116 Paris
prenom.nom@talan.com,
https://www.talan.com/global/fr

Résumé. ATLAS (**A**utomated **T**opological **L**ink **A**nalysis **S**ystem) est un outil conçu comme une extension web pour automatiser le maillage interne des sites en s'appuyant sur des techniques avancées d'intelligence artificielle. Se basant sur des techniques d'embeddings modernes, ATLAS analyse le contenu de chaque page pour générer des représentations sémantiques précises. Grâce à un score de similarité, l'outil détermine automatiquement quelles pages doivent être reliées, favorisant une structure de liens conforme aux principes du cocon sémantique. Cette approche vise à optimiser la navigation interne et à maximiser la visibilité du site sur les moteurs de recherche, tout en renforçant l'expérience utilisateur. Ainsi, ATLAS aide les spécialistes du SEO en simplifiant la gestion des liens internes et en améliorant le référencement naturel. ATLAS s'inspire du concept de cocon sémantique pour la création des liens.

1 Introduction

Dans le contexte numérique actuel, un bon référencement naturel également appelé Search Engine Optimisation (SEO) est crucial pour assurer la visibilité des sites web. Le maillage interne, structure de liens entre les différentes pages d'un site, joue un rôle fondamental dans ce référencement mais reste une tâche complexe réservée aux experts. Des études récentes (Zhang et Cabage, 2017; Sharma et al., 2019) soulignent l'importance des liens internes pour l'indexation des contenus, tandis que les révélations sur l'algorithme de Google (Goodwin, 2024) confirment leur rôle central dans l'évaluation de la pertinence d'un site.

Les techniques d'embedding en NLP pour traiter la similarité sémantique ont considérablement évolué. Des modèles contextuels comme BERT (Kenton et Toutanova, 2019) et son extension pour la gestion de texte (Reimers, 2019), cohabitent désormais avec des modèles plus légers et optimisés tels que MiniLM (Wang et al., 2020). Une autre approche dérivée des modèles d'IA génératifs est l'usage des embedders comme Ada (OpenAI, 2022).

Bien que l'idée d'utiliser des embeddings pour la clusterisation de texte dans d'autres circonstances ait déjà été traitée (par exemple dans Goel (2024)), nous n'avons pas, à notre connaissance, d'usage de cette approche pour la création automatique et intelligente de liens internes d'un site web. En revanche, il existe des outils tels que Screaming Frog (2010) ou

développés directement par Google (2006), mais ces derniers servent surtout à la visualisation de l'état courant d'un site, les améliorations et corrections étant à la charge de l'utilisateur.

ATLAS propose une approche novatrice basée sur l'intelligence artificielle pour automatiser la création d'un maillage interne optimisé. L'outil exploite des représentations sémantiques calculées via embeddings pour générer des liens pertinents entre les pages. Cet article présente son architecture, ses fonctionnalités et évalue son efficacité à travers plusieurs métriques, comparant ses résultats avec des maillages réalisés manuellement.

2 Présentation de l'outil

L'outil ATLAS se présente actuellement pour des raisons historiques comme une extension TYPO3 publique dont une version est déjà disponible [1]. Cette extension propose une première implémentation bout à bout de notre cas d'étude, partant d'un site web et proposant des suggestions pour améliorer son maillage interne. Nous souhaitons automatiser au maximum la création d'un maillage conforme à l'approche du *cocon sémantique* (Twaino, 2022; Interactive, 2023), une notion qui apparaît de plus en plus dans la littérature SEO. Cette notion est une méthode de structuration des contenus web qui relie de manière logique et thématique les pages d'un site, afin d'améliorer la compréhension par les moteurs de recherche et de renforcer le référencement.

En parallèle du développement de l'extension, nous travaillons également sur une version Python (disponible sur git [2]) afin de réaliser une étude sur le meilleur choix d'embedder via plusieurs métriques que nous avons mises en place. Cette version Python servira de base pour une future API qui permettra d'élargir son usage, hors du cadre limité d'une extension TYPO3.

2.1 Fonctionnement général

L'idée principale de l'outil est de partir d'un site web en ligne, puis de recalculer un maillage interne automatique en évaluant la proximité sémantique des pages de ce site afin de savoir si deux pages doivent être reliées ou non. Conformément au schéma 1, nous avons les étapes clés suivantes (le bloc **Perspectives** sera traité ultérieurement) :

1. **Entrée :** On sélectionne un site web à analyser par l'outil.
2. **AnalyseSEO :** L'intégralité des contenus est stockée et, pour chaque page, il faut désormais calculer un vecteur d'embedding correspondant. Ce composant permet aussi d'avoir une visualisation du maillage existant du site étudié.
3. **Vectorisation :** Pour chaque page, un vecteur est généré en utilisant une des solutions que nous avons présélectionnées : MiniLM, Ada2, Text-embedding-3-small et CamemBERT. Une fois les vecteurs calculés, ils sont stockés dans une base de données vectorielle FAISS (Facebook AI Similarity Search) (Johnson et al., 2019).
4. **Génération du maillage :** Un maillage est généré en calculant la similarité entre les vecteurs, deux à deux. Si la similarité dépasse un seuil prédéfini, un lien (bidirectionnel) est suggéré et intégéré au maillage généré.

1. Lien : `https://extensions.typo3.org/extension/semantic_suggestion`
2. Lien : `https://github.com/friteuseb/seoanalyse`

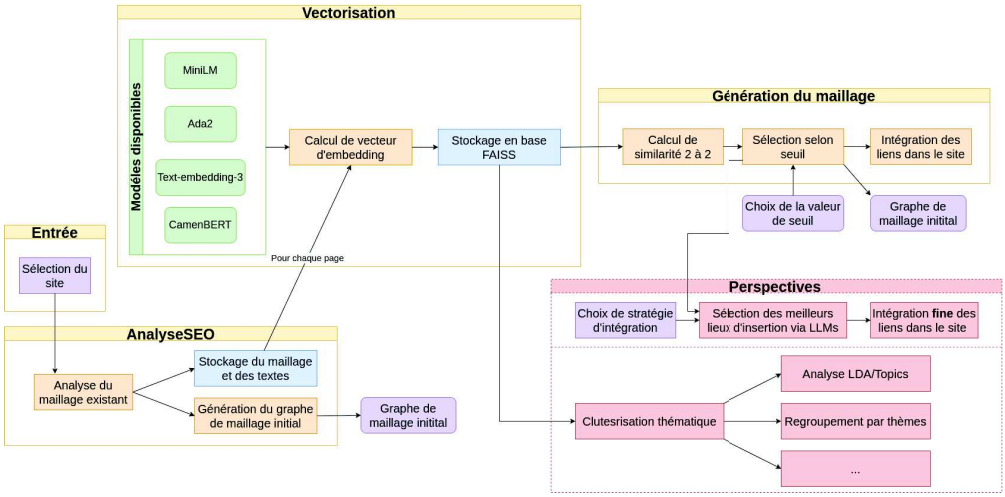


FIG. 1 – *Fontionnement d'ATLAS, du site analysé au maillage généré*

2.2 Modèles d'embedding utilisés

— **MiniLM** : Modèle léger de Microsoft, optimisé pour les tâches de similarité textuelle, permettant un bon compromis entre performance et rapidité.
— **Ada2** : Développé par OpenAI, spécialisé dans la compréhension sémantique profonde avec une excellente performance sur les nuances contextuelles.
— **Text-embedding-3-small** : Version compacte du modèle text-embedding-3 d'OpenAI, combinant efficacité et précision.
— **CamemBERT**(Martin et al., 2019) : Adaptation française de BERT, particulièrement performante sur les contenus en français grâce à son pré-entraînement sur un large corpus francophone.

2.3 Justification de l'approche par similarité

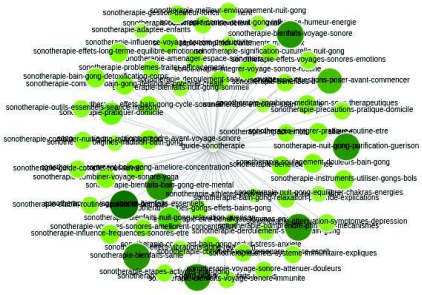
La création de liens entre pages similaires, même si elles partagent une page parente commune, présente plusieurs avantages clés :

— **Navigation contextuelle** : Les utilisateurs naviguant sur une page spécifique sont souvent intéressés par des contenus similaires. Relier directement ces pages permet de réduire le nombre de clics nécessaires, ce qui améliore l'expérience utilisateur.
— **Renforcement thématique** : Les moteurs de recherche utilisent les liens internes pour comprendre la structure thématique d'un site. Des liens directs entre contenus similaires renforcent cette compréhension, au-delà de la simple hiérarchie parent-enfant.
— **Distribution du "jus de lien"** : Dans le cadre du SEO, les liens directs permettent une meilleure distribution du PageRank entre pages thématiquement liées, renforçant leur autorité collective sur le sujet et permettant d'améliorer la robustesse de la navigation, de renforcer les associations thématiques et de faciliter la découverte de contenus connexes.

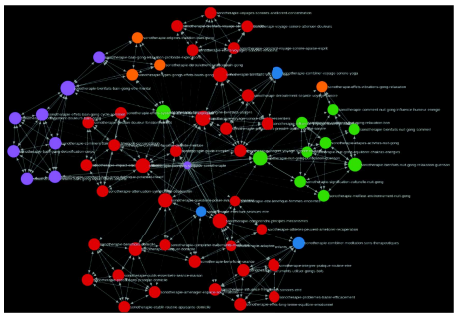
2.4 Visualisation du maillage

De nombreux outils sont disponibles pour accompagner les experts SEO dans l'analyse d'un site, notamment en matière de visualisation. Ces outils sont essentiels pour une lecture claire de la structure d'un site. De la même manière, un outil de visualisation propre à ATLAS a été développé pour permettre une meilleure comparaison de nos résultats.

Notre solution de visualisation se distingue des outils existants comme *Screaming Frog* en excluant volontairement les liens liés aux menus, évitant ainsi des connexions non pertinentes entre les pages hiérarchiques, conformément au concept de cocon sémantique. Elle intègre également une clusterisation par thème des pages afin d'améliorer la lisibilité générale du cocon.



(a) Représentation graphique typique des outils SEO du marché, les menus de navigation faussent la représentation, le menu étant central et connecté à toutes les pages.



(b) Représentation graphique d'ATLAS, ignorant les menus de navigation. Les différentes thématiques du site sont segmentées automatiquement à travers un code couleur.

FIG. 2 – *Comparaison de graphes avec et sans le menu d'un site.*

3 Evaluation d'ATLAS

3.1 Protocole expérimental

L'évaluation d'ATLAS repose sur deux cas d'étude judicieusement choisis. Le site "Échecs et Tennis" (50 pages), avec ses thématiques distinctes et sa structure non optimisée, permet d'évaluer la capacité d'ATLAS à distinguer et organiser des contenus hétérogènes. Le cocon sémantique "Sonothérapie" (+60 pages), optimisé par un expert, sert de référence pour mesurer la capacité d'ATLAS à reproduire un maillage professionnel sur une thématique unique.

L'analyse utilise quatre modèles d'embedding (MiniLM, Ada2, Text-embedding-3-small et CamemBERT), calculant la similarité via FAISS avec un seuil de 0.60 (0.20 pour MiniLM) selon la formule :

$$S(q,p) = \frac{1}{1 + d_{L2}(E(q), E(p))} \tag{1}$$

où d_{L2} est la distance euclidienne entre les embeddings $E(q)$ et $E(p)$ des pages q et p. Cette formule normalise la distance euclidienne dans l'intervalle [0,1], où 1 indique une similarité parfaite et 0 une dissimilarité totale. Les seuils sont ajustés selon les caractéristiques de chaque modèle, MiniLM nécessitant un seuil plus bas (0.20) du fait de sa tendance à produire des scores de similarité plus faibles.

3.2 Métriques d'évaluation

Pour évaluer la qualité du maillage généré par ATLAS, nous avons sélectionné six métriques complémentaires :

— **Pages orphelines :** Nombre de pages isolées du maillage.
— **Densité :** Ratio entre les liens existants et les liens possibles.
— **Accessibilité en 3 clics :** Proportion des pages accessibles depuis n'importe quelle autre page en moins de 3 interactions.
— **PageRank :** Calcul du PageRank (Gleich, 2015) en fonction du maillage.
— **Recouvrement des maillages :** Pour chaque paire de maillages (y compris l'initial), nous avons calculé la proportion de liens identiques.
— **Nombre d'ilots :** Nombre de groupes de pages complètement isolés.

Ces métriques couvrent la connectivité globale (densité, accessibilité), la qualité structurelle (pages orphelines, îlots) et la pertinence sémantique (PageRank, recouvrement) du maillage, s'inspirant des bonnes pratiques SEO tout en intégrant des indicateurs spécifiques à notre approche.

4 Résultats et analyse

4.1 Impact sur site non optimisé

Le cas "Echecs et Tennis" démontre la capacité d'ATLAS à transformer significativement un site non optimisé :

Modèle	Nb Liens	Page Orphelines (%)	Densité (%)	Accessibilité	PageRank moy	Nb Ilots
cocon-t0	18	50.0	0.04% (+0.0%)	13.0% (+0.0%)	4.10	5
ada2-t60	180	0	0.47% (+1231.6%)	100.0% (+666.7%)	4.19	1
camembert-t60	159	0	0.42% (+1076.2%)	90.0% (+590.0%)	4.08	1
3-small-t60	28	15.8	0.07% (+107.1%)	20.0% (+53.3%)	4.20	7
minilm-t20	20	31.6	0.05% (+48.0%)	15.0% (+15.0%)	4.06	6

TAB. 1 – *Etude sur le cas "échecs et tennis" avec un maillage interne faible*

— **Structuration automatique** : Réduction des pages orphelines de 50% à 0% avec Ada2, démontrant l'efficacité de la détection thématique automatique.
— **Densification pertinente** : Augmentation du maillage de 18 à 180 liens tout en maintenant un PageRank stable (4.10 à 4.20), validant la pertinence des liens créés.

— **Amélioration de l'accessibilité** : Passage de 13% à 100% d'accessibilité en 3 clics, optimisant l'expérience utilisateur.
— **Performance des modèles gratuits** : CamemBERT atteint des scores proches d'Ada2 avec 90% d'accessibilité et 0% de pages orphelines.

4.2 Comparaison avec un maillage expert

L'analyse du site "Sonothérapie" valide la capacité d'ATLAS à reproduire un maillage professionnel :

— **Préservation structurelle** : Maintien d'une densité optimale (0.12% à 0.15%) et d'une excellente accessibilité (>86.9%).
— **Stabilité qualitative** : PageRank moyen stable entre 5.42 et 5.67 (vs 5.76 initial), confirmant la pertinence sémantique.
— **Reproduction experte** : Les modèles d'embedders plus récents (Ada-2, CamenBert) permettent au cocon généré d'être plus proche de celui de l'expert SEO.
— **Similarité entre trois modèles performants :** Les modèles **Ada2-t60**, **CamemBERT-t60**, et **3-small-t60** affichent des performances très proches. Tous trois maintiennent une densité de 0.15%, une accessibilité élevée (entre 88.5% et 91.8%), et un PageRank moyen stable (5.65 à 5.67 pour Ada2-t60 et 3-small-t60, 5.42 pour CamemBERT-t60).

Modèle	Nb Liens	Page Orphelines (%)	Densité (%)	Accessibilité	PageRank moy	Nb Ilots
cocon-t0	485	0.0	0.12	100.0	5.76	1
ada2-t60	558	3.3	0.15	91.8	5.67	2
3-small-t60	557	3.3	0.15	88.5	5.65	1
camembert-t60	550	3.3	0.15	90.2	5.42	1
minilm-t20	498	10.0	0.14	86.9	5.39	1

TAB. 2 – *Comparaison avec le cocon sémantique "sonothérapie" – état de l'art*

4.3 Comparaison des modèles

Le Tableau 3 présente la matrice de recouvrement entre les différents maillages générés, permettant d'évaluer la cohérence entre les modèles et leur alignement avec le maillage expert initial (cocon-t0).

	cocon-t0	ada2-t60	3-small-t60	camembert-t60	minilm-t20
cocon-t0	-	48.2	28.9	49.9	35.7
ada2-t60	48.2	-	44.2	81.1	61.2
3-small-t60	28.9	44.2	-	44.7	38
camembert-t60	49.9	81.1	44.7	-	62.9
minilm-t20	35.7	61.2	38	62.9	-

TAB. 3 – *Matrice de recouvrement entre les différents maillages (%)*

L'étude révèle des performances différenciées selon les modèles :

— **Ada2** : Performance optimale avec un recouvrement de 48.2% avec le maillage expert et une forte cohérence avec CamemBERT (81.1% de recouvrement).
— **CamemBERT** : Alternative gratuite performante, démontrant un recouvrement de 49.9% avec le maillage expert, et une forte similarité avec les choix d'Ada2.
— **MiniLM** : Solution légère offrant un bon compromis coût/performance avec 35.7% de recouvrement avec l'expert, nécessitant un seuil adapté (0.20).
— **Text-embedding-3-small** : Performance plus modérée avec 28.9% de recouvrement avec l'expert, mais maintenant une cohérence acceptable avec les autres modèles.

5 Conclusion et perspectives

ATLAS démontre sa capacité à s'adapter à différents contextes en transformant significativement les sites non optimisés tout en reproduisant fidèlement les choix experts sur des sites déjà structurés. L'efficacité des modèles gratuits (MiniLM, CamemBERT) rivalisant avec les solutions payantes suggère une démocratisation possible des bonnes pratiques SEO, rendant l'optimisation du maillage interne accessible à un plus large public.

Notre approche par analyse sémantique automatique soulève une hypothèse prometteuse : les liens générés par analyse du contenu réel pourraient s'avérer plus pertinents que ceux créés selon des règles hiérarchiques rigides. Cette innovation ouvre la voie à plusieurs axes de développement :

— **Validation empirique** : Études A/B sur le comportement utilisateur, analyse des métriques d'engagement (taux de rebond, temps de session), impact sur le référencement naturel
— **Enrichissement intelligent** : Insertion contextuelle des liens, génération d'ancres optimisées, suggestions de contenu complémentaire
— **Adaptation dynamique** : Mise à jour automatique du maillage lors de modifications de contenu, détection des opportunités de création de contenu
— **Internationalisation** : Support multilingue, prise en compte des spécificités culturelles dans l'analyse sémantique

Ces développements, combinés à l'expérience acquise sur les premiers cas d'usage, visent à transformer ATLAS en une solution complète de gestion du maillage interne. Au-delà de la simple automatisation, l'objectif est de créer un outil capable d'accompagner intelligemment la stratégie éditoriale des sites web, en suggérant non seulement des liens pertinents mais aussi des opportunités d'amélioration du contenu.

Références

Gleich, D. F. (2015). Pagerank beyond the web. *siam REVIEW 57*(3), 321–363.

Goel, R. (2024). Using text embedding models and vector databases as text classifiers with the example of medical data. *arXiv preprint arXiv :2402.16886*.

Goodwin, D. (2024). HUGE Google Search document leak reveals inner workings of ranking algorithm. `https://searchengineland.com/google-search-document-leak-ranking-442617`. Consulté le 20/11/2024.

Google (2006). Google search console. https://search.google.com/search-console. Consulté le 20/11/2024.

Interactive, L. (2023). Cocon sémantique seo : définition et guide de création. Consulté le 20/11/2024.

Johnson, J., M. Douze, et H. Jégou (2019). Billion-scale similarity search with gpus. *IEEE Transactions on Big Data 7*(3), 535–547.

Kenton, J. D. M.-W. C. et L. K. Toutanova (2019). Bert : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, Volume 1, pp. 2. Minneapolis, Minnesota.

Martin, L., B. Muller, P. J. O. Suárez, Y. Dupont, L. Romary, É. V. de La Clergerie, D. Seddah, et B. Sagot (2019). Camembert : a tasty french language model. *arXiv preprint arXiv :1911.03894*.

OpenAI (2022). Introducing text-embedding-ada-002. https://openai.com/blog/introducing-text-and-code-embeddings. Consulté le 20/11/2024.

Reimers, N. (2019). Sentence-bert : Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv :1908.10084*.

Screaming Frog (2010). Screaming frog seo spider tool. https://www.screamingfrog.co.uk/seo-spider/. Consulté le 20/11/2024.

Sharma, D., R. Shukla, A. K. Giri, et S. Kumar (2019). A brief review on search engine optimization. In *2019 9th international conference on cloud computing, data science & engineering (confluence)*, pp. 687–692. IEEE.

Twaino (2022). Cocon sémantique : Définition, exemples et guide de création pour booster votre seo. Consulté le 20/11/2024.

Wang, W., F. Wei, L. Dong, H. Bao, N. Yang, et M. Zhou (2020). Minilm : Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems 33*, 5776–5788.

Zhang, S. et N. Cabage (2017). Search engine optimization : Comparison of link building and social sharing. *Journal of Computer Information Systems 57*(2), 148–159.

Summary

ATLAS (**A**utomated **T**opological **L**ink **A**nalysis **S**ystem) is a tool designed as a web extension to automate internal linking for websites, leveraging advanced artificial intelligence techniques. Using modern embeddings approaches, ATLAS analyzes the content of each page to generate precise semantic representations. Through a similarity score, the tool automatically determines which pages should be linked, promoting a link structure that adheres to the principles of semantic cocooning. This approach aims to optimize internal navigation and maximize the site's visibility on search engines, while also enhancing the user experience. As such, ATLAS provides a turnkey solution for SEO specialists seeking to simplify internal link management and improve organic search rankings. ATLAS draws inspiration from the concept of semantic cocooning for link creation.