AJC EAM

Academic Journal on Computing, Engineering and Applied Mathematics

2021
Volume 2 Issue 2

Academic Journal on Computing, Engineering and Applied Mathematics

Image Designed by xb100 / Freepik

UNIVERSIDADE FEDERAL DO TOCANTINS

ISSN: 2675-3588

Universidade Federal do Tocantins

Dados Internacionais de Catalogação na Publicação (CIP)

Academic Journal on Computing, Engineering and Applied Mathematics (AJCEAM) [recurso eletrônico] / Universidade Federal do Tocantins, Curso de Ciência da Computação. – vol. 02, n. 02 ([abril/agosto], 2021) – Palmas - TO, UFT, 2021. ISSN nº 2675-3588.

Quadrimestral no primeiro ano de publicação 2020
Semestral.
Disponível em:
https://sistemas.uft.edu.br/periodicos/index.php/AJCEAM/index

1. Ciência da Computação - periódico. 2. Matemática Aplicada. 3. Computação Aplicada. 4. Engenharias. 5. Ciências Exatas. I. Universidade Federal do Tocantins.

CDD 22.ed. 004

Ficha Catalográfica elaborada por Edson de Sousa Oliveira – CRB/2 – 1069.

Expediente

Realização
Fundação Universidade Federal do Tocantins (UFT)
Quadra 109 Norte, Avenida NS-15, ALCNO-14 | Bloco III | sala 214 |Plano Diretor Norte | 77001-090 | Palmas / TO | Brasil

Periodicidade
Este periódico possui periodicidade semestral e utiliza a Licença Creative Commons 4.0 - CC BY-NC 4.0. Contudo, a publicação dos artigos em modalidade avançada ou ahead of print, ou seja, tão logo os manuscritos aprovados sejam editados para publicação, é possível. O AJCEAM não possui taxas de publicação, tanto pouco de submissão de manuscritos, sendo totalmente gratuita para autores e leitores.

Indexadores
Google Acadêmico, desde 9 de maio de 2020
International Standard Serial Number – ISSN, desde 28 de maio de 2020
Crossref, desde 7 de junho de 2020
Revistas de Livre Acesso – LivRe, desde 24 de junho de 2020

Sumário

Editorial (Português): Academic Journal on Computing, Engineering and Applied Mathematics

Warley Gramacho da Silva[1]

[1] *Universidade Federal do Tocantins, Palmas / TO, Brasil*

Inicio esse editorial parabenizando a equipe AJCEAM, na pessoa do seu Editor-chefe Rafael Lima de Carvalho, pelo esforço dedicado a realização e manutenção dessa revista. Este número traz dois trabalhos, que pela excelente abordagem científica, contribuirá no processo de disseminação da revista, pois são trabalhos com grande potencial de divulgação.

O primeiro trabalho, intitulado "*Desenvolvimento de um bot baseado em IA para o MMORPG Tibia*", escrito por Sousa e Carvalho [1], apresenta o desenvolvimento e avaliação de desempenho de um bot para jogar o MMORPG Tibia de uma forma automatizada enquanto mantém o comportamento humano. Os autores alcançaram esse resultado por meio da combinação de técnicas de Inteligência Artificial e algoritmo de pesquisa em grafo A*. O bot desenvolvido foi capaz de alcançar resultados competitivos quando comparado a jogadores humanos.

O segundo trabalho, Silva e Carvalho [2], através de seu artigo intitulado "*Using Latent Semantic Indexing as a metric for evaluating research potentialities through Innovation Public Policies*", propõe o uso de Indexação de Latência Semântica (do inglês, *Latent Semantic Indexing - LSI*), uma técnica de busca e recuperação da informação, para associar projetos de pesquisa de determinada instituição com as políticas de inovação do governo federal. O estudo apresenta o potencial da utilização do LSI para classificação e indexação dos projetos e detectar seu potencial de inovação. Uma solução que pode auxiliar as instituições na atuação frente à política nacional de inovação.

Por fim, agradecer aos autores em [1, 2] pela escolha do AJCEAM para terem seus trabalhos divulgados. Assim, desejo boa leitura e excelente aprendizado.

Trabalho duro é inútil
para aqueles que não acreditam
em si mesmos.
Naruto Uzumaki

Referências

[1] T. Castanheira Retes de Sousa and R. L. d. Carvalho, "Desenvolvimento de um bot baseado em ia para o mmorpg tibia," *Academic Journal on Computing, Engineering and Applied Mathematics*, vol. 2, no. 2, p. 1–9, ago. 2021. [Online]. Available: https://sistemas.uft.edu.br/periodicos/index.php/AJCEAM/article/view/12529

[2] R. Oliveira Silva and R. Lima de Carvalho, "Using latent semantic indexing as a metric for evaluating research potentialities through innovation public policies," *Academic Journal on Computing, Engineering and Applied Mathematics*, vol. 2, no. 2, p. 10–15, ago. 2021. [Online]. Available: https://sistemas.uft.edu.br/periodicos/index.php/AJCEAM/article/view/12528

Development of an intelligent agent for Tibia MMORPG

Thiago Castanheira Retes[1] and Rafael Lima de Carvalho[1]

[1] *Federal University of Tocantins, Computer Science Department, Tocantins, Brazil*

Reception date of the manuscript: 26/06/2021
Acceptance date of the manuscript: 05/08/2021
Publication date: 11/08/2021

Abstract— Artificial Intelligence has always been used in designing of automated agents for playing games such as *Chess*, *Go*, *Defense of the Ancients 2*, *Snake Game*, *billiard* and many others. In this work, we present the development and performance evaluation of a reactive agent for the RPG Game Tibia. The intelligent agent is built using a combination of AI techniques such as graph search algorithm A* and computer vision tools like template matching. Using four algorithms to get global position of player in game, handle its health and mana, target monsters and walk through the game, we managed to develop a fully automated Tibia agent based in raw input image. We evaluated the performance of the agent in three different scenarios doing ten sessions of fifteen minutes and five sessions of one hour, collecting and analyzing metrics such as *XP Gain*, *Supplies Usage* and *Balance*. The simulation results show that the developed agent is able to play the game consistently according to in-game metrics.

Keywords— Automated bots, TIBIA, Artificial Intelligence, Computer Vision

I. Introduction

Gaming is a huge industry that generated $22.4 billion in sales in 2014, according to the Entertainment Software Association [1]. In a Massively Multiplayer Online Role-Playing Game (MMORPG), thousands of players can play together inside a persistent game world [2]. According to Cano[1], "of the tens of millions of players who play games daily, 20 percent play massively multiplayer online role-playing games (MMORPGs)". Inside the the MMORPG platforms, a lot of players use to trade virtual goods within thriving in-game economies.

According to Ferreira et al.[3], Tibia is a Massive Multiplayer Online Role-Playing Game (MMORPG), developed by CipSoft. Tibia is also one of the oldest games in the genre (1997), with a community that surpasses 500.000 players, for which 40% of these are Brazilian.

Artificial Intelligence has many fields of actuation when dealing with games. After many research efforts in order to come up with an AI that is able to play against a human being, in 1997 the Deep Blue computer, made by IBM, was able to beat the world chess champion Garry Kasparov[4]. Another milestone happened in 2009 when an AI named Fuego beat Chou Chun-hsun, the world champion in GO, a challenging board game far more complex than chess in number of possibilities (even though it was using a small board than the official one)[4] . Moreover, in March 2016 Google's AlphaGo AI, defeated Lee Sedol, the winner of 18 international titles, in March 2016 [5].

Looking to survey approaches of solving Snake Game and build an AI Bot that solves it in minimal amount of steps, Sharma et al. [6], compared Best First Search, A* Search, A* Search with forward checking, Random Move and Almighty Move according to its traits and proposes an AI bot mixing them, after the evaluation of them, they proposed an AI bot that used Best First Search in the first 4 fruits, then A* with Forward Checking for the next 34 fruits and then Almighty Move for the next 62 fruits, which resulted in an AI Bot that can solve Snake Game in 100 iterations, which outperforms the other algorithms alone. The authors propose in the discussion session a different game to train normal players using a shadow snake controlled by their AI Bot which the player has to mimic their movements to earn score.

Looking at the field of electronic games, the programmers use AI mostly to model automated bots that compose the behavior of enemies as well as game assistants. On the other hand, there is a field of actuation in AI that tries to create an agent to play on behalf of a human player in electronic games. These piece of software are known as *game bots*[7]. Therefore, game bots usually uses AI in order to play the games automatically and this function is considered a great challenge and a complex computational benchmark.

Towards the construction of such game bots, Dharmawan and Hanafiah [8] discuss the usability of a clicker bot based in template matching for *gacha games* which are games with chances in freemium business models, these kind of games share a trait with MMORPGs which are the common need to farm lots of resources, that the bots saves the user lots of time and effort by automating the farming. In its work, the author prefer to present the results according to the computational

Contact data: Thiago C. Retes, thiagoretes2@gmail.com

performance and accuracy of the functions implemented in the clicker bot, in this paper we took a different approach, evaluating the performance of agent according to some measures given by the MMORPG itself.

Regarding pathfinding in video games, Cui and Shi[9] discuss about A* algorithm and its related optimizations Hierachical Pathfinding A*(HPA*) and Navigation Mesh(NavMesh). Moreover, the authors describe optimizations to the heuristic function used by A* by using an overestimating function in order to trade shortest path for a faster search, then it describes ways to optimize memory usage of A* with approaches such as allocating a minimum node bank in memory and using Iterative Deepening A*(IDA*). The authors also suggest using hash tables for the list of closed nodes and a binary heap for the list of open nodes in order to optimize the data structures used by A*. Cui and Shi[9] also discuss relevant applications in the game industry presenting how well known titles such as Age of Empires II, Civilization V, and World of Warcraft handle the pathfinding problems.

In order to develop an autonomous navigation system for a robot, Ayala-Raggi *et al.* [10] used two view-based approaches, one using SURF, RANSAC and Proscutes analysis and another one using template matching with Normalized Cross Correlation(NCC), both using a panoramic image composed of 3 images of the camera in a robot that covers 76º x 3 = 228º. The main results indicated that the SURF approach was more consistent in regard to the addition of obstacles which NCC template matching failed.

Since publication of Mnih et al.[11], that used raw input of screen to feed a deep learning model for playing 49 Atari games and achieved super-human levels in 29 of them, academic community mainly focused in solving problems adopting raw image input by applying machine learning techniques. On the other hand, in this paper we try to bring another light in these problems by using knowledge-based systems with AI and Computer Vision techniques to play a complex game using raw image as the input for the proposed agent, but not using expensive computational techniques such as deep learning.

The main objective of this work is to present the development of a reactive agent, based on artificial intelligence and image processing, that is capable of playing Tibia using visual information and generating common inputs. The proposed agent is then evaluated in three different scenarios and the game metrics are collected in order to measure its performance during the game play.

The present work is organized as follows: Section II presents the necessary background with the two main techniques used to develop the intelligent agent, such as A* algorithm and Template Matching for dealing with the raw image input. Section III explains the game Tibia, such as the map size, interface and mechanics. Section IV presents the methodology used to create the agent and evaluate it. Section V presents the simulations we did for each scenario presented in Section IV the results and our interpretation of the results. Section VI presents our conclusions about the work and our possible future developments.

II. BACKGROUND

Since the proposed intelligent agent is based in two main techniques, in Subsection a it is shown the basic background of A* search algorithm. In Subsection b shows the background knowledge about Template Matching, the computer vision algorithm employed as well.

a. A Algorithm*

The A* algorithm is a path-searching and graph traversal algorithm, which is usually used due to its completeness and optimal efficiency[12]. It can be seen as an extension to Dijkstra's algorithm. Its main difference to Dijkstra's algorithm is the cost function which is: $f(x) = g(x) + h(x)$ where $g(x)$ is the cost to reach current node and $h(x)$ can be a graph of costs or an heuristic function that estimates the cost to the goal node.

For its optimality it is necessary that the heuristic function $h(x)$ meets certain requirements. The first condition is that $h(x)$ has to be an admissible heuristic. An admissible heuristic means that it never overestimates the cost to reach the goal. Since $g(x)$ is the cost of the actual state, then $f(x) = g(x) + h(x)$ never overestimates the cost along of each path[12].

Another requirement is that the heuristic function $h(x)$ has to be consistent. Consistency means for every state x and every successor x', the cost of reaching $h(x)$ has to be lower than the cost of the action from x to x' plus $h(x')$, which can be seen in Equation 1 [12]:

$$h(x) <= c(x,a,x') + h(x') \tag{1}$$

According to Hart et al.[13], given a node s, a set T of goal nodes and the successor operator Γ, the A* search algorithm can be described as:

1. Mark s as open and calculate $f(s)$
2. Select the open node n whose value of $f(n)$ is the smallest, resolving ties arbitrarily, but prioritizing any node $n \in T$
3. If $n \in T$, mark n as closed and terminate the algorithm.
4. Otherwise, mark n as closed, apply the successor operator Γ to n. Calculate $f(n')$ for each successor of n' and mark as open, every successor that has not already been marked as closed yet. Remark as open any closed node n' which is a successor of n and for which $f(n')$ is smaller now than it was when n' was marked as closed. Go to Step 2.

b. Template Matching

Template matching is a method used in digital image processing in order to find a small image in a larger one. It can be implemented by simply sliding the small image window in the larger image as a 2D convolution and comparing patch of larger image under the template one. Several comparison methods can be used, such as Cross Correlation or Square Difference.

The problem of template matching has been studied extensively in the field of computer vision for years. Many

approaches were applied in almost all tasks of computer vision such as stereovision, camera calibration, recognition of objects, etc[14].

In Brunelli[15], the author presents the simplest template matching technique used in computer vision known as planar distribution of light. This technique takes the intensity values and transform them into a vector x, that can be compared, in a coordinate-wise fashion. This produces a spatially congruent light distribution represented in a similar form by vector y. The formula is given by Equations 2 and 3:

$$d(x,y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2 = \frac{1}{N}||x-y||_2^2 \tag{2}$$

$$s(x,y) = \frac{1}{1+d(x,y)} \tag{3}$$

The lower the value of $d(x,y)$ or the higher the value of $s(x,y)$, the better in indication of pattern similarity.

Since $s(x,y)$ is the comparison function, there are numerous functions that can be used in its place. The library we used in our intelligent agent was OpenCV 4.2.0[16], and it implements three of them with their normalized variants. Our agent was coded to use the comparator named correlation coefficient normalized.

III. Tibia

As in any MMORPG, the main idea of the game is to evolve the player's character through hunting. This action gives the player money and XP (indicator for experience), and allows it to participate in the game history through quests and wars.

So an intelligent agent is a tool that has been used for a long time by players to evolve their characters through endless hunts. The main idea behind an intelligent agent is to automate an action to evolve the player's character through simple algorithms. In this work the goal is to evaluate an agent throughout measures of efficiency in order to produce the maximum XP per hour on Tibia. The map of Tibia is composed of 2304x2048 SQM(Square Metre) and 14 floors which gives a total exploration size of 66.060.288 SQMs, whether all parts of the map is available to walk through.

As it can be seen in Fig. 1, the Graphics User Interface in Tibia is kept as simple as possible. At the upper right corner, there is the *minimap* which shows the player's position in the game and enables it to navigate through the world with a click. Immediately beneath, there is located the current equipped items. Below it, there are the health and mana bars, which indicates the current amount of life and mana of the character. At the center of the screen, there is the main game screen. It consists of a matrix of 15x11 SQM and shows the actions of the characters in the player vision. Beneath it, there is the hotkeys bar, which contains the shortcuts to spells, items and text. Immediately bellow it, there are the text channels which shows all the actions that happens during the game play. The rest of the interface is placeholder for containers and windows that show specific statistics.

Tibia is a MMORPG which every action has some kind of cooldown involved. In general, there are 4 (four) types of cooldowns that are important to consider when building an intelligent agent for Tibia:

- *Movement*: It is the cooldown between movement from an SQM to another. Every SQM has a type that determines the amount of cooldown in conjunction with the player level.
- *Attack Spell Group*: every attack spell adds some global attack spell cooldown when cast, in general has a cost of 2 (two) seconds from the last time the spells were used;
- *Healing Spell Group*: Every healing spell adds in general a 1 (one) second global healing spell cooldown;
- *General Item use*: Usually the player can use one item for every second with some exceptions like the potion of *mana* shield;
- *Basic Attack*: The basic attack is based in weapon equipped and has a cooldown of 2 seconds.

a. Vocations

Tibia has 4 vocations, each of them with unique traits and its advantages and disadvantages.

- *Knight*: is the vocation which excels at melee combat. It has the highest amount of health points gaining 15 health points, 25 oz of capacity and 5 mana points per level. It evolves faster in melee skills(club, axe and sword) and shielding skill. Its spells are mostly focused in melee combat and tanking damage. Its main damage attribute is based on the weapon it uses(some weapons can have elemental damage thus making Knights elemental damage dealers).
- *Paladin*: is the distance combat class. It has a balanced amount of health points and mana points gaining 10 health points, 20 oz of capacity and 15 mana points per level. It evolves faster in distance fighting and shielding skill. Its spells are mostly focused in distance combat and self regeneration.
- *Druid*: is the healing mage class. It has the least amount of health points and the most amount of mana points gaining 5 health points, 10 oz of capacity and 30 mana points per level. It evolves faster in magic level. Its spells are mostly focused in healing, although it has some nice combat spells with Terra and Ice attributes.
- *Sorcerer*: is the damage dealer mage class. It has the least amount of health points and most amount of mana points, exactly like druids, sorcerer gains 5 health points, 10 oz of capacity and 30 mana points per level. It evolves faster in magic level. Its spells are most focused in dealing damage with Fire and Energy attributes.

b. Skills

Each player in Tibia can evolve different skills during their gameplay, they represent their expertise with said type of action, these skills can be separated in 3 main types:

Fig. 1: Tibia Interface. (a) In Red: Minimap, (b) In Yellow: Equipped Items, (c) In Orange: Health and Mana Bars, (d) In Green: Main Game Screen, (e) In Blue: Shortcuts and Hotkeys, (f) In Violet: Text Channels

- *Melee skills*: They are 3 skills (club, axe, sword), responsible for the damage done using melee weapons and knight attack spells. Can be evolved by using basic attack while using a weapon of the skill type.
- *Distance skill*: It is the skill responsible for the damage done using distance weapons (bows and crossbows). Can be evolved by using basic attack while using a bow or a crossbow.
- *Magic Level*: It is the skill responsible for the power of spells in general(healing and attack spells). Can be evolved by spending mana points.

c. Combat

The action of hunting creatures in Tibia involves attacking them until their health points reach zero. Players can attack creatures in two ways: basic attack and attack spells. Most creatures that give experience points in Tibia are aggressive, this means that they will engage in combat with the player as soon its with a certain range of them. Creatures don't have a global spell cooldown which means every two seconds they can cast all of their spells, however its unlikely to happen because the spells are cast by chance. In order to survive the engagement players can use healing spells and potions to restore their health points and to keep using spells they can use mana potions to restore their mana points.

d. Hunting

A variety of places in Tibia has spawns of creatures which each of them gives a specific amount of experience points and items dropped(loot) while having a specific amount of health points and attack(which is a combination of basic attack and spells), some creatures are better to hunt for XP and others for loot. What makes a good hunting spot in Tibia is the combination of amount, density, respawn time of creatures in it, the difficult of hunting there and its popularity. Since Tibia is not a instanced MMORPG and its world is shared between the entire game world, the best hunting spots are usually disputed which may turn them a bad choice for hunting, so the popularity of a hunting spot is a factor to consider too because most hunting spots does not support more than a player hunting which will decrease the amount creatures each player will be able to kill. The hunting in Tibia basically consists of killing creatures doing a specific circular route, ensuring that when the player completes the route, monsters have respawned. Each vocation has particular hunting spots that are best suited to it.

IV. Methodology

The world of Tibia is composed of 2304x2048 SQMs and 14 floors, which gives a total size of 66.060.288 SQMs. This amount of information must be loaded by the proposed agent in order to use the GPS (Global Positioning System) of Tibia. Since each SQM is represented by 3 bytes (BGR), the whole map size accounts to 189 MB.

In order to make a fully automated hunt agent we built 4 agents that run simultaneously:

- GPS Algorithm: it is responsible for retrieving the global position of the character;
- Healing Algorithm: it is responsible for healing the character and not letting it die;
- Target Algorithm: it is responsible for targeting monsters, hence defeated them and "looting" (collect) their items;
- Cave Bot Algorithm: it acts to walk around the hunting place according to the GPS;

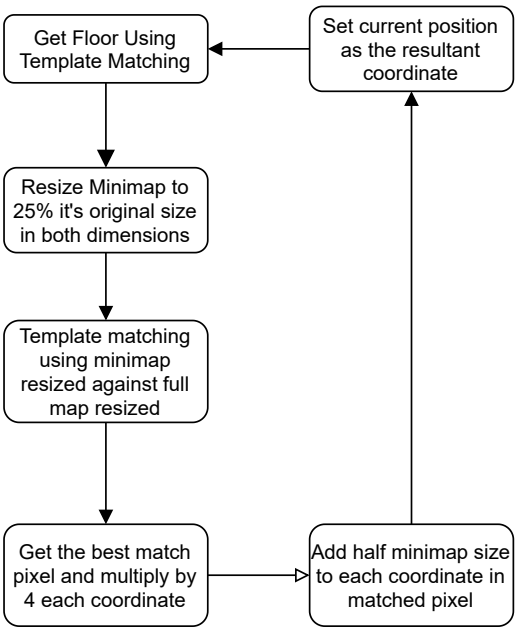

Fig. 2: Diagram of GPS Algorithm

In Fig. 2 is shown the proposed GPS algorithm, which is an optimized template matching to get the minimap and match against the full map. The first step is getting the floor which will be a template matching against each of 14 possible floor positions located at the right side of the minimap which provides the correct floor. After that, to find a good match, the algorithm resizes the minimap to 25% of its origianl size, at each dimension and match against a copy of the full map, also resized to 25% of its original size. This process reduces the search space to 1.179.648 pixels, which in BGR means 3.375MB.

The healing algorithm is a simple pixel check in the health and mana bars. If the specific pixel has not the expected color, then a configured healing shortcut is sent to the game.

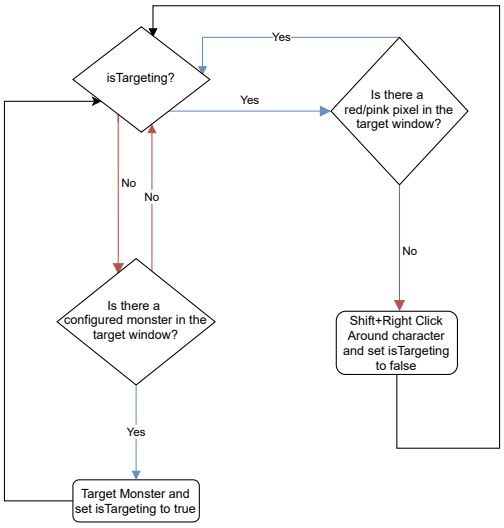

Fig. 3: Diagram of Targeting Algorithm

In Fig. 3 is shown the target system, which is a template matching to get the monsters in the target window and matches them against a preconfigured monsters hunting list. If the match is above some preconfigured threshold and there is no red/pink pixel in the target window (which means the character has already engaged some other monster), then the algorithm will click the monster to attack it. After defeating the engaged monster, the target algorithm sends specific keybindings to the positions around the character, in order to collect the items generated by the victory.

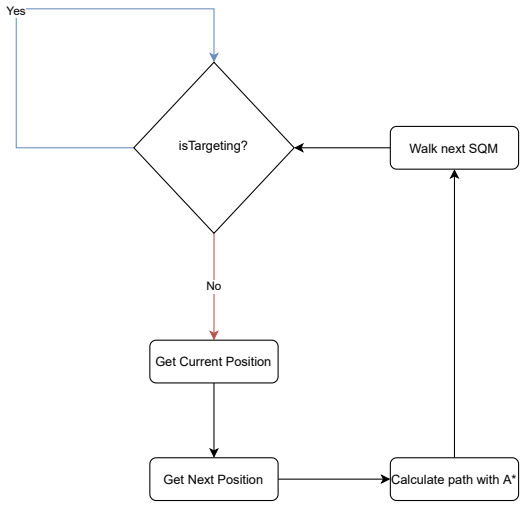

Fig. 4: Diagram of Cavebot Algorithm

In Fig. 4 is shown the cave bot algorithm, which takes as input a circular buffer of positions to walk. It walks to the next position while there is no monster targeted. If there is a targeted monster it stops, then continue walking. The walking stage is based on the A* search of map space doing optimal search decision based on the speed of the current position towards destination. The map of tibia contains information about the speed at each SQM which is used as weights for A* algorithm.

a. Agent Evaluation

During the gameplay, the TIBIA's character has some performance measures. In general, TIBIA measures the following indexes:

1. *XP Gain*: this is the mainstream measure. It increases when the player kills creatures and decreases when the player dies;
2. *Loot*: this is the sum of the value of all items dropped by creatures in gold pieces;
3. *Supplies*:this is the sum of the value of all items wasted to hunt the creatures in gold pieces;
4. *Balance*: this is the Loot minus the Supplies in gold pieces;
5. *Killed Monsters*: the amount of killed monsters.

Because Killed Monsters and Loot are linear dependencies of the other three, during the simulations, we evaluate the agent performance using three main measures: *XP Gain*, *Supplies*, and *Balance*.

V. Simulations and Results

There is a variety of monsters in Tibia for hunting although since we only had a character of level 89 we had a limited

variety of hunting places to choose, for this reason we chosen three hunting places that would fit our test character.

In order to evaluate the performance of the proposed agent, we designed three test scenarios: Scenario I: Wasp cave(Ab'Dendriel), Scenario II: Nomads Cave(Ankrahmun/Darashia), and Scenario III: Lion Sanctuary(Darashia). The scenarios are described in Table 1. For each scenario, we ran the agent for 10 times, with sessions of 15 minutes and for 5 times, with sessions of 1 hour. For each run session, the agent began in similar starter points.

Our choice of doing 10 sessions of 15 minutes was due to two main reasons. First to test its reliability and stability against bigger sessions in terms of results and second due to Tibia stamina system, which gives bonus XP in the first 180 minutes of gameplay each day, hence, with our choice we have been able to conduct one test scenario per day.

TABLE 1: DESCRIPTION OF THE EVALUATED SCENARIOS.

Scenario	Description
I. Wasps	This is the easy level and the main monsters the agent has to face is Wasps. Each Wasp may attack only for a low amount of damage which is healed by only food the agent is eating. The main difficulties an automated agent can face here are the map exploration.
II. Nomads	This is mid-level hunting spot which holds a high-demand item(rope belt) which has a good value for Loot hunting. The main difficulties an automated agent may face here are the high amount of creatures.
III. Roaring Lions	This is a high-level popular hunt spot for great XP Gain with little to no waste in terms of Balance. The main difficulties an automated agent may face here are the high amount of damage creatures inflict to it.

a. 15 minutes sessions

In Fig. 5 it is shown the agent performance during the 10 sessions, using the Scenario I. It can be noticed that the balance variable has a high standard deviation, deviating from the average especially in runs 2,6,7,8,9. This happened because of the random nature of drops in Tibia. Once a Wasp monster is defeated, the dropped items does not follow a pattern. Notice, however, that the XP gain is stable, because it is related to the experience in defeating the related monster. Considering the *Supllies* measure, it can be seen that there was a higher use of supplies in run 1 and a suddenly loss during run 5. This happens due to the use of *brown mushrooms* as supplies which is used once every 264 seconds. So as the agent is running during sessions of 15 minutes, it means that it will be used around $15 * 60/264 = 3.4$ mushrooms for every session. Therefore, there are sessions that will use more than the expected 3 mushrooms.

In Fig. 6 it is shown the agent performance during 10 sessions, using Scenario II as well. It can be noticed a lower standard variation in balance when compared with Scenario I. This is due to the drop rate of valuable items from nomads being higher than of the wasps. This summed to the higher

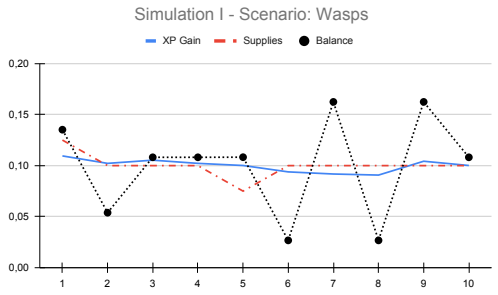

Fig. 5: Normalized results of Wasps Scenario.

amount of monsters defeated resulted in a lower standard deviation of the balance variable. The variation that can be noticed in the XP Gain is due the nature of the cave scenario. Thus, sometimes the agent end up walking through a hole in the cave, having restarted its route.

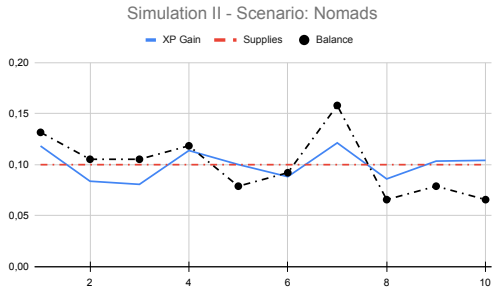

Fig. 6: Normalized results of Nomads Scenario.

In Fig. 7 it is shown the agent performance during 10 sessions, using Scenario III. It can be noticed a high standard deviation in balance. This is due to the low drop rate of items from Roaring Lions. It can be noticed that the XP Gain and Supplies are tied due to the amount of supplies it takes to defeat a Roaring Lion. This is the unique Scenario that the proposed agent had to use *Mana Potions* to defeat the monsters, since each *Mana Potion* costs 56 gold pieces the balance got hit by it. It can be noticed too, that there is a standard deviation in XP too, which is mainly because some players killed monsters of the hunting place while we were running the session during some sessions and secondarily because the agent actually killed more Roaring Lions during the session sometimes.

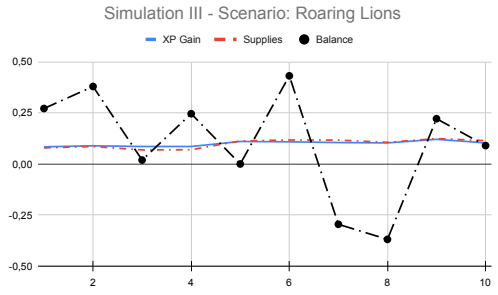

Fig. 7: Normalized results of Roaring Lions Scenario.

We also conducted another analysis, plotting each measure into the three evaluated scenarios. In Fig. 8, it is shown the XP for all scenarios (with no normalization). Looking at the

8, as expected the XP Gain of Roaring Lions was clearly above Wasps and Nomads due to the nature of each scenario. We note too that nomads has a high XP Gain which is of course due to the Scenario being a higher level of difficult than Wasps, since a Wasp gives 36 XP while a Nomad gives 90 XP.

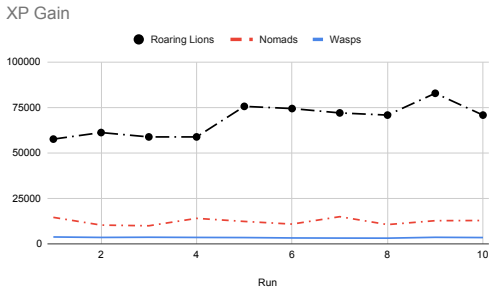

Fig. 8: XP Gain results of all simulations.

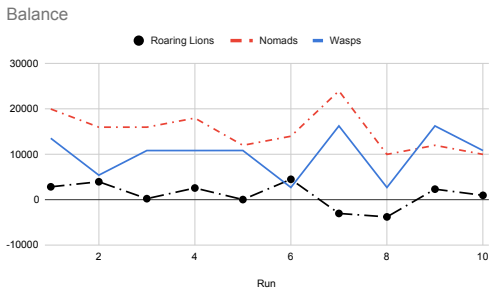

Fig. 9: Balance results of all simulations.

In fig. 9 as expected balance of Roaring Lions scenario runs were the lowest since its nature is basically XP Gain without any profit. We also noticed that Wasps balance are under Nomads balance because Nomads scenario is a higher level of difficulty than Wasps. In addition, it is noticeable that there is a high standard deviation in the balance which is due to the randomness nature of loot in Tibia.

b. 1 hour sessions

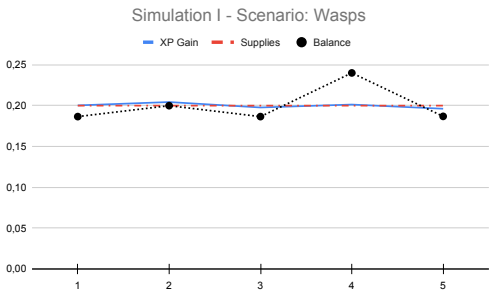

Fig. 10: Normalized results of wasps for 1 hour session.

In fig. 10 is shown the results of 5 sessions using the Scenario I. It can be noticed a much lower standard deviation in comparison with fig. 5, which is probably due to the higher duration of the sessions. This is a good result since it signals the stability of the agent performance in this scenario. The Balance still has higher deviation due to its random nature but the impact has been smaller in these higher duration sessions.

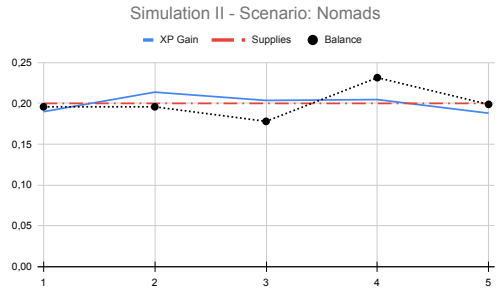

Fig. 11: Normalized results of Nomads for 1 hour session.

In fig. 11 is shown the results of 5 sessions using the Scenario II. It can be noticed a lower standard deviation in comparison with fig. 6, which shows us that the higher the duration of the sessions, stabler are our agent results. The Balance metric still shows a some standard deviation, but lower when compared to the 15 minutes sessions.

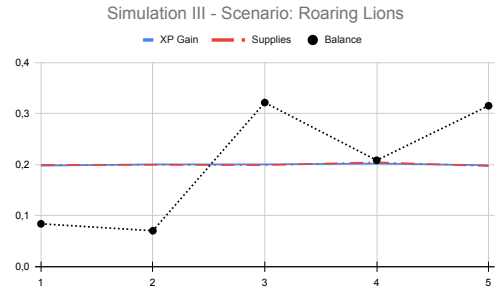

Fig. 12: Normalized results of Roaring Lions for 1 hour session.

In fig. 12 is shown the results of 5 sessions using the Scenario III. It can be noticed that there is almost no deviation in XP Gain and Supplies, which signals the stability of the proposed agent in this scenario. The Balance still has a high deviation due to the low drop rate of some high value items from Roaring Lions.

c. Comparison with human players

In order to evaluate the performance of the proposed agent we got a volunteer to run a session of 1 hour in Scenario I and Scenario II. The volunteer was not able to run a session in Scenario III, after trying to go to the hunting place one time and dying, he gave up due to his skills not being enough for it. Our volunteer was a 26 years old with basic knowledge of the game. The sessions were done using the same character which we conducted the agent sessions.

In fig. 13 is shown the comparison of the session done by the volunteer with the best and worst sessions done by the agent in Scenario I. It can be noted that the agent achieved almost the same XP Gain of the human player, which means it killed almost the same amount of creatures as the human player did. In regards to balance, due to the random nature of loot, the agent may had a bad luck in its worst session but almost the same balance of the volunteer in its best session. The Supplies were in the range of 60 to 140, so it did not appear on the chart.

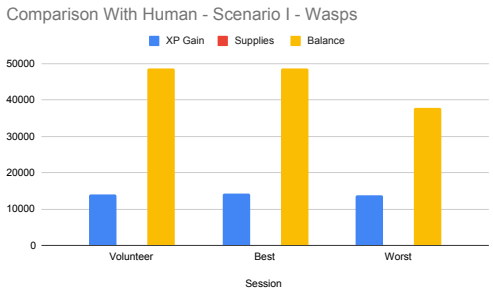

Fig. 13: Comparison of Human Volunteer with Best and Worst session of 1 hour for Scenario I - Wasps.

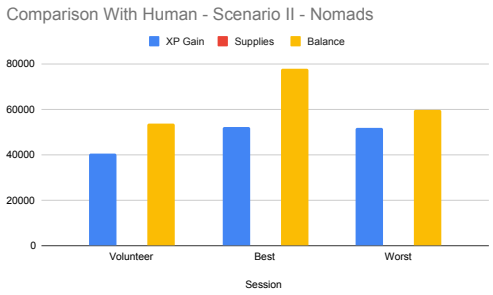

Fig. 14: Comparison of Human Volunteer with Best and Worst session of 1 hour for Scenario II - Nomads.

In fig. 14 is shown the comparison of the session done by the volunteer with the best and worst sessions done by the agent in Scenario II. It can be noted that the agent's performance was better than the volunteer in both XP Gain and Balance. Supplies were in the range of 130 to 140, so they did not appear on the chart.

When doing a preliminary comparison with public sessions recorded by some gamers, in the environment described here as Nomads Scenario the proposed agent produced around 15.000 balance of gold pieces per session of 15 minutes which translates into 60.000 gold pieces per hour against 78.000 of Cyf [17] and in the environment described here as Roaring Lions Scenario the proposed agent produced a result of around 70.000 experience points per session of 15 minutes which translates into 280.000 experience points per hour against 350.000 experience points per hour of [18]. This results are primarily due to faster input/output of the agent and its nature as an automated agent of never being affected by emotions.

VI. Final Remarks

This paper presented the development and evaluation of a reactive agent for the MMORPG Tibia. The proposed agent used only raw data input as the main screen of the game and an algorithm based on Template Matching to process the images. Furthermore, it has been implemented a GPS module based on the A* algorithm which is used to guide the agent through the map and target monsters along the way.

The evaluation consisted of three distinct scenarios with three different levels of difficulties. For each scenario, the agent is submitted in sessions of 15min and 1 hour, and three main measures are collected: XP, Supplies, and Balance. As expected, in the most difficult scenario, the Roaring Lions, the agent made XP close to what an human player would do, but without dying a single time in the evaluated sessions. In the easier scenarios, suited to farm gold pieces, the agent behaved well farming values near to what a human player would farm.

Future research investigations include: a) improvement of the proposed agent framework to allow the evaluation of players, gathering all their inputs and outputs visible through the screen, in order to evaluate the agent's performance against other players; b) comparison with agents using other approaches such as reinforcement learning; and c) benchmark with more human volunteers; d) Research the optimum value of map resize for the best performance of the GPS Algorithm.

References

[1] N. Cano, *Game hacking: developing autonomous bots for online games*. No Starch Press, 2016.

[2] J. Y. Wang, "Combat State-Aware Interest Management for Massively Multiplayer Online Games," Master's thesis, TECHNISCHE UNIVERSITÄT MÜNCHEN, 2017.

[3] N. Ferreira, P. Trovo, and S. Nesteriuk, "Emergence in game design: Theoretical aspects and project's potentialities," in *International Conference on Distributed, Ambient, and Pervasive Interactions*. Springer, 2017, pp. 597–611.

[4] J. Schaeffer, M. Müller, and A. Kishimoto, "Go-bot, go," *IEEE Spectrum*, vol. 51, no. 7, pp. 48–53, 2014, cited By 1.

[5] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton *et al.*, "Mastering the game of go without human knowledge," *nature*, vol. 550, no. 7676, pp. 354–359, 2017.

[6] S. Sharma, S. Mishra, N. Deodhar, A. Katageri, and P. Sagar, "Solving the classic snake game using ai," in *2019 IEEE Pune Section International Conference (PuneCon)*, 2019, pp. 1–4.

[7] A. Kang, S. Jeong, A. Mohaisen, and H. Kim, "Multimodal game bot detection using user behavioral characteristics," *SpringerPlus*, vol. 5, 2016.

[8] T. Dharmawan and N. Hanafiah, "Clicker bot for gacha games using image recognition," *Procedia Computer Science*, vol. 179, pp. 598–605, 2021, 5th International Conference on Computer Science and Computational Intelligence 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050921000521

[9] X. Cui and H. Shi, "A*-based pathfinding in modern computer games," *International Journal of Computer Science and Network Security*, vol. 11, no. 1, pp. 125–130, 2011.

[10] S. E. Ayala-Raggi, P. d. J. González, S. Sánchez-Urrieta, and A. Barreto-Flores, "A simple view-based software architecture for an autonomous robot navigation system," in *Image Analysis and Recognition*, M. Kamel and A. Campilho, Eds. Cham: Springer International Publishing, 2015, pp. 287–296.

[11] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015. [Online]. Available: http://dx.doi.org/10.1038/nature14236

[12] S. J. Russell and P. Norvig, *Artificial Intelligence: a modern approach*, 3rd ed. Pearson, 2009.

[13] P. E. Hart, N. J. Nilsson, and B. Raphael, "A formal basis for the heuristic determination of minimum cost paths," *IEEE Transactions on Systems Science and Cybernetics*, vol. 4, no. 2, pp. 100–107, July 1968.

[14] B. Cyganek, *Object Detection and Recognition in Digital Images*. John Wiley and Sons Ltd, 2013.

[15] R. Brunelli, *Template matching techniques in computer vision: theory and practice*. Wiley, 2009. [Online]. Available: http://gen.lib.rus.ec/book/index.php?md5=3F5932473E6E9C09092322AC4F5EC6E0

[16] G. Bradski, “The OpenCV Library,” *Dr. Dobb’s Journal of Software Tools*, 2000.

[17] C. Plays, “Rope belt farm (level 20+ / 70k+ profit) - bonus : Arito’s task quest.” 2018. [Online]. Available: https://www.youtube.com/watch?v=f_si_oDb1LI

[18] Itexo, “Ek 54 lion’s rock - 350k/h - os melhores lugares de hunt para knights,” 2018. [Online]. Available: https://www.youtube.com/watch?v=7q6fD6q2cYc

Using Latent Semantic Indexing as a metric for evaluating research potentialities through Innovation Public Policies

Renan Oliveira Silva[1] and Rafael Lima de Carvalho[1]

[1] *Universidade Federal do Tocantins, Computer Science Department, Tocantins, Brazil*

Reception date of the manuscript: 26/06/2021
Acceptance date of the manuscript: 02/08/2021
Publication date: 10/08/2021

Abstract— Public innovation policies usually define strategies for public research organizations, such as universities, in order to guide the next research projects of such organizations. Sometimes, it is difficult to know the actual state of an organization when a new policy is released by the government. The objective of this paper is to present the application of Latent Semantic Analysis, a technique of information retrieval, in order to create an index and automatically classify research projects, using text fields like title and abstract, to areas and subareas defined by related terms. It is also proposed a case study of about 200 projects from five graduate programs of the *Universidade Federal do Tocantins*. The proposed solution was capable of satisfactorily classify each project to the areas and subareas of a recent policy from the Science, Technology, Innovations, and Communications Ministry. In this way, the university could have some decision-making information, and the results could sustain for which internal policies could be implemented to maximize its actuation faced to the national innovation policy.

Keywords—Latent Semantic Analysis, Science and Technology, Research and Innovation Policies.

I. Introduction

Throughout history, humanity has been continuously evolving as a society. We have been learning over the centuries to form the concept of our "modern society". This concept, when applied acceptably, can give certain rights to the people and should develop laws or policies attempting to provide services to the social well-being. Such policies, both public and private, need to be evaluated because the implementation of these policies relies on investments. In the public context, a policy is drafted aiming to address a solution to a certain public problem. A policy is commonly associated with a program and there are expectations that it will fix a problem, which implies that there is something to be fixed, i.e., a problem [1].

The policies evaluation can occur in different moments within the policy cycle, each one is used for different purposes, according to when it is performed [2]:

- ***ex-ante***: it happens at the beginning of the policy creation process, though, it provides strategical information that determines the continuation of a policy.
- **interim**: it lies sometime between the *ex ante* and the *ex post* evaluations. It can help avoid or correct mistakes within the development process. Moreover, it assesses the results of the implementation phase, providing control information.
- ***ex post***: it is performed at the end of the policy cycle, verifying the policy impacts. Furthermore, it may help in design and decision-making on similar projects in the future.

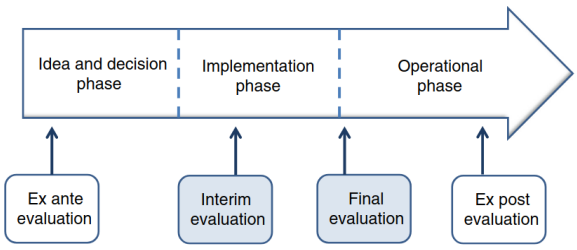

Fig. 1: The policy evaluation in different moments.

The Figure 1 [2] illustrates the policy cycle evaluation timeline, highlighting the policy evaluation in different moments. According to [3], matching the requirements of policymakers with the skills and experience of evaluators can reveal crucial divergences in perspectives. Likewise, such divergences may affect the delivery and implementation of evaluation studies that serve the desired policy purposes. Additionally, the policies arise, interact and have effects in a messy, complex, multi-level, and multi-actor reality [4]. The Figure 2 [3] illustrates the complexity of defining a metric for the evaluation study.

Contact data: Renan Oliveira Silva, renan.oliveira@mail.uft.edu.br

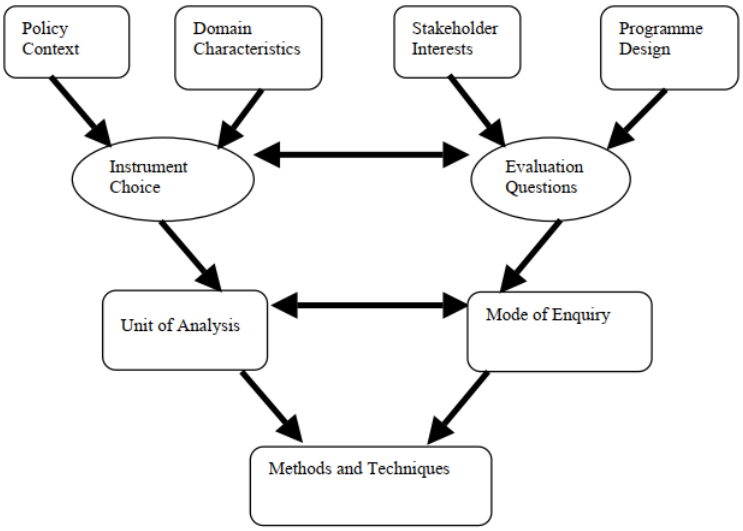

Fig. 2: A framework for formulating an appropriate evaluation methodology.

In Brazil, most of the public policies concerning research and development are provided by the MCTIC (*Ministério da Ciência, Tecnologia, Inovações e Comunicações*[1]). In 2020, MCTIC released an ordinance defining priority areas with regard to research projects in national scope, for the development of technologies and innovations for the period from 2020 to 2023 [5]. In such document, they have pointed five great strategic areas known as: Strategical Technologies, Skilled Technologies, Production Technologies, Technologies for Sustainable Development and Technologies for Quality of Life. Each area is divided into more specific subareas, defining the strategies for the development of further research projects.

Many Brazilian organizations are responsible for the development of research projects. Much of such production comes from public universities [6], which have many professors actuating as researchers. Usually, these researchers register their projects using the Lattes Platform[2] (a national platform used to store curriculum data for researchers). So, to know which projects the researchers are working on in a period, the Lattes data can be used, but textual information should be processed to give a semantic evaluation of such data. Therefore, it could be time-consuming to evaluate the potentialities of such research organizations, when an innovation policy is released.

In this work, we propose the use of Latent Semantic Indexing to automatically associate projects of a research organization into an innovation policy formed by areas and subareas, with free terms classifying each subarea. The research question is: given the representation of research, such as documents consisting of title and abstract, and an innovation policy, which describes terms of areas and subareas, is it possible to use LSI to automatically classify each project into the policy to evaluate the current potential of innovation of such organization?

This document is organized as follows: Section II describes the necessary background covering the main technical subjects such as LSI and correlated textual tools. In Section III we discuss, through some examples, the LSI effectiveness in different applications. In addition, in Section IV we present the adopted methodology, which defines the sample database, scope and delimitation, as well as the main pipeline connecting the textual tools. Further, Section V presents the analysis of the results when applying the methodology to a Case Study of a University and the Brazilian Science and Technology ministry's innovation policy for the country. Finally, Section VI shows some final remarks of this research as well as some future insights.

[1]Science, Technology, Innovations and Communications Ministry
[2]http://lattes.cnpq.br/

II. Background

In this paper, we use Latent Semantic Indexing (LSI) as a policy evaluation metric. The LSI, also referred to as Latent Semantic Analysis (LSA), is a technique for information retrieval that takes advantage of implicit structures and recognizes semantic relations between terms and documents [7]. This approach also produces more concise results in comparison with matching term approaches.

a. Singular value decomposition

The core of the LSI is the Singular Value Decomposition (SVD). As the name implies, the SVD decomposes the term-document matrix, which will be discussed later, and transforms it into a lower-dimensional matrix. At this point, the purpose is to deal with non-square and probably non-symmetric matrices [8]. Applying the SVD, the term-document matrix A_{txd} will be decomposed into the product of T_{txn}, S_{nxn}, and D_{dxn}:

$$A_{txd} = T_{txn} S_{nxn} (D_{dxn})^T$$

where t is the number of terms, d the number of documents and $n = \min(t, d)$ [9]. The columns of T and D are orthonormal, it means that $TT^T = D^T D = I$. Furthermore, the S is a diagonal matrix in the descending order containing the eigenvalues of $A_{txd}(A_{txd})^T$ that are the same for $(A_{txd})^T A_{txd}$. In the SVD, the eigenvalues are known as *single values*. The restriction of T, S and D to their first $k < n$ rows, results in

the T_{txk}, S_{kxk} and D_{dxk} matrices [9]. The product of these matrices is $\hat{A}$, which is:

$$\hat{A}_{txk} = T_{txk} S_{kxk} (D_{dxk})^T$$

and it is the best square approximation of A by a matrix of rank k as defined in the equation $\Delta = ||A - \hat{A}||$ [9]. We are not going to discuss SVD deeply in this paper, but a brief description is that the decomposition allows us to represent a huge matrix (term-document matrix) in a lower dimension. That is why $k < n$ and the $\hat{A}$ matrix has a lower rank than A. Additionally, the $\hat{A}$ matrix represents the synonyms more precisely than A [9].

b. Preprocessing

Before applying the LSI in the input data, it is necessary to take out unwanted parts of the dataset in a process known as preprocessing. At the beginning of this process, the stopwords are removed. The stopwords show up frequently in natural language documents, but it provides minimal contextual information [10]. Example stopwords include each, about, such, and the. The next step is to reduce words with different grammatical inflections. In such a process, there are two main approaches: stemming and lemmatization. Both techniques reduce words into a base form, but they work in a slightly different manner.

The stemming consists of converting morphological forms of words to their stem, a stem does not need to be an existing word in the dictionary, but all its variants should map to this form after the stemming process [11]. On the other hand, lemmatization transforms a word into a lemma: a canonical form of a lexeme. The lexeme is a set of all word inflections that have the same meaning. The stem is formed by plenty of rules used in the stemmer algorithm, while the lemma is chosen by convention to represent a certain lexeme.

c. Dictionary and bag of words

After the preprocessing, all the documents within the dataset are analyzed and the documents' words are stored in a dictionary. That dictionary is a lookup table containing the document frequency of a term as well as where in the postings file the per-document counts can be accessed [12]. The postings file is an inverted file with the per-document counts and the IDs of the documents, as well as the position of each term occurrence.

Thus, the bag of words (BoW) is created and, according to [13], it is a simple representation of text that is used in retrieval and classification models. In this representation, a document is considered to be an unordered collection of words with neither syntactic nor statistical relationships between them. Then, all documents are "reshaped" according to the BoW, i.e., the documents are represented by the term occurrences. At this point, the purpose is to arrange the documents as vectors in the vector space model.

d. Document-term matrix and vector space model

According to [13], in the vector space model, documents and queries are assumed to be part of a t-dimensional vector space, where t is the number of index terms and a document is represented by a vector of index terms as follows:

$$D_i = (d_{i1}, d_{i2}, ..., d_{it}),$$

A document collection containing n documents (document-term matrix) of term weights, each row representing a document and each column describing the weights assigned to a term for a certain document, can be represented like this:

	$Term_1$	$Term_2$	...	$Term_t$
$Document_1$	d_{11}	d_{12}	...	d_{1t}
$Document_2$	d_{21}	d_{22}	...	d_{2t}
⋮	⋮	⋮		⋮
$Document_n$	d_{n1}	d_{n2}	...	d_{nt}

Because the documents are vectors to be arranged in a vector space [12], the Figure 3 illustrates a vector space with two coordinates (terms).

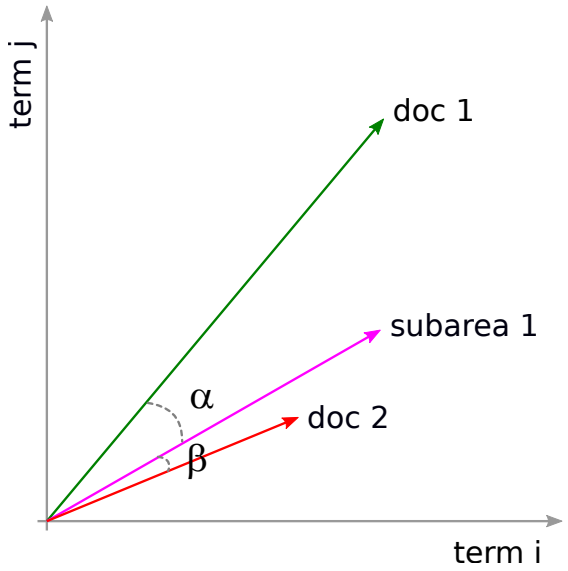

Fig. 3: Documents plotted in a vector space with two coordinates.

As vectors plotted in a vector space, the documents can be compared to determine the similarities between them. Both query and documents are vectors plotted in the vector space, consequently, they are pointed and there are angles among them [12]. One approach broadly used to compute such angles is the cosine similarity. The documents are ranked by computing the similarity between the query and each document in the document collection.

When plotting the documents in the vector space, the model ranks the documents according to a number of topics, which varies depending on the dataset. A common practice is to compute the ideal number of topics for each dataset, given an interval. For example, if the interval is $[1, 100]$, an algorithm will calculate the topic coherence for each number within this interval. The number of topics with the greatest coherence score will be used in the model to rank the dataset. After the ranking, the model utilizes a threshold given by the user, if a document score is less than the threshold, it will not be shown in the ranking results.

e. TF-IDF

As previously mentioned, the vectors are pointed and because their coordinates are terms, each coordinate in a vector is a weight that points the vector. At first, a weight represents the presence of terms in a document, because of that, the vectors assume two values: zero or one. When a term is present in a document, it is assigned one to the term coordinate, otherwise, it is assigned zero. But considering just the presence or absence of a word is a naive approach, that becomes clear when realizing that the presence of a term in a document does not tell us anything about the documents context [14].

A more precise approach is assigning weights as term per-document counts, which gives us more refined results. But to improve the ranking of the documents even more, we need to consider some constraints. Using the term per-document count directly as the vector weight may be accurate enough, mainly if we consider very frequent words in a document. That is why, documents with high term frequencies can be ranked over documents with a more interesting context, for example. Another point is that very frequent words throughout the collection do not have the same relevance as rare words when distinguishing a document context.

Aiming to address a solution for the problems described above, the TF-IDF is a commonly used function for assigning the vector weights. The TF-IDF consists of both term frequency (TF) and inverse document frequency (IDF), while the TF value considers directly the term occurrence, the IDF penalizes very common terms throughout the document collection and rewards rare words [12].

III. Related Works

Since its formulation in the 1980s [15], several researches have applied LSI. A well-known application is in search engines, which uses the LSI power to classify a massive amount of stored documents, which became globally used after the internet break out in the 1990s. The search engines utilize information retrieval techniques, such as LSI, to rank millions of documents according to a user's query.

Besides the query-search usage, the LSI can be applied for different purposes. Some researches like [16], [17], and [18] utilize it to approach different problems. Even if [19] research aims at a different objective in comparison to ours, both apply the LSI in a quite similar manner, classifying papers according to title and abstract. Moreover, some works combined different techniques to enhance the LSI power, as seen in [20] and [21].

IV. Methodology

a. Summary

Primarily, it was necessary to create a dataset to be used when applying the LSI. A research was made by [22] and the data refers to the works of the UFT researchers and the data sources was both MCTIC (*Ministério da Ciência, Tecnologia, Inovações e Comunicações*) and the *Currículo Lattes*. Then, the research results were used to mount a dataset which is used as input by this work.

At this point, the given input is refined in the preprocessing and, lately, plotted in the vector space model. As discussed previously, the dataset documents (papers) and the queries (pseudo-documents) are vectors in the vector space. Now, the model computes the similarity between each query and the documents according to a threshold and a number of topics. Lastly, the ranking results are plotted in a graph.

b. Scope and delimitation

The first data source is composed of two main documents: a) "ENCTI 2016-2020", which is a document with the Brazilian's national strategy for Science, Technology and Innovation for the time interval between 2016 and 2020; and b) MCTIC Ordinance No. 1,122/2020 (later partially amended by the MCTIC Ordinance No. 1,329/2020), which defines the priority research areas of the national strategy, covering the period from 2020 to 2023. Each area is separated into sub-areas or sectors, namely:

- **Strategical Technologies**: Spatial; Nuclear; Cybernetics; and Border and Public Safety;
- **Skilled Technologies**: Artificial Intelligence; Internet of Things; Advanced Materials; Biotechnology; e Nanotechnology;
- **Production Technologies**: Industry; Agribusiness; Communications; Infrastructure; and Services;
- **Technologies for Sustainable Development**: Smart Cities; Renewable Energies; Bioeconomy; Solid Waste Treatment and Recycling; Pollution Treatment; Monitoring, prevention and recovery from natural disasters and environmental; and Environmental Preservation; and
- **Technologies for Quality of Life**: Health; Sanitation; Water Security; and Assistive Technologies.

This research is applied to a Case Study of the Federal University of Tocantins (UFT). Currently, the UFT has 33 graduate programs. Based on the aforementioned areas, this research considered a sample composed of the following graduate programs: *Biotecnologia*, *Agroenergia*, BEC, CTA, MCS and Profnit.

Such a sample includes the name of the researchers and their projects within the graduate programs. All the information about the projects was collected and there are 95 researchers in this sample. To take the most recent data, it was considered only active projects from 2018 to 2020. The projects' data source was the *Currículo Lattes*, which is a governmental platform used to update researchers' curricula. From that point of view, it has been collected an amount of 211 projects, which contains title and abstract.

The pseudo-documents used are collections of terms representing each subarea of the MCTIC. The terms composing the collections are common terms shared by projects located in the same subarea. The pseudo-documents were composed by [22] and it was not considered any constraints when eliciting such terms, the only concern was selecting terms that were contextually important to the subarea and could, as better as possible, describe it with the source of [23].

c. Flowchart

The flowchart in Figure 4 demonstrates the whole evaluation process and, as already stated, a dataset is a collection of pseudo-documents characterizing the MCTIC's subareas. Therefore, the input is prepared to be analyzed and ranked by the LSI, which computes the similarity between queries and documents. After the entire process, the results are shown in a graph.

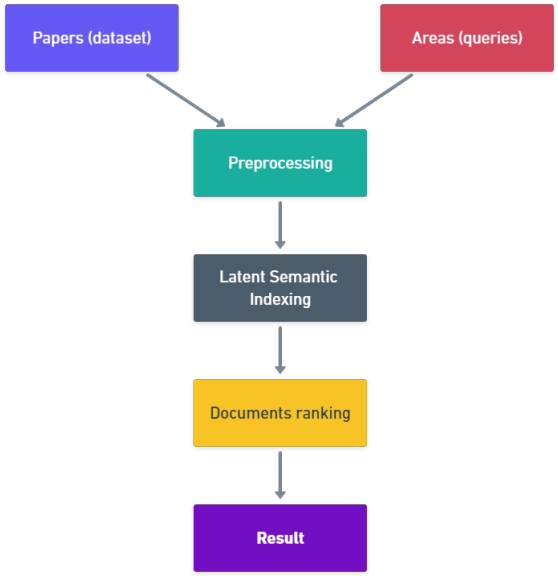

Fig. 4: The evaluation process flowchart.

V. Results

The evaluation process is based on analyzing the similarity of each project in the dataset with each given subarea. Each subarea of the MCTIC belongs to a great area, for that reason, we classify a document in an area through one of its subareas. The benefit of matching subareas is that we can provide a deeper and more detailed evaluation. The Figure 5 illustrates the subarea matching.

Document	Subarea 1	Subarea 2	...	...	Subarea N	Matched Subarea
Doc_1	0.366016	0.802408	...	...	0.566351	Subarea 2
Doc_2	0.373501	0.598308	...	...	0.665918	Subarea N
...	...	...	...	..	...	...
...	...	...	...	...	...	...
Doc_N	0.713722	0.142225	..	...	0.211479	Subarea 1

Fig. 5: Document ranking for each subarea.

If a document matches a subarea of a great area, we consider it as being part of that particular great area, for example: if a document matches Sanitation, we consider it as being part of Quality of Life. Looking at Figure 6, it can be seen that the majority of projects were semantically associated with the areas of Production and Sustainable Development. The balance of the number of projects in the areas Skilled Technologies, Quality of Life and a little less associated with Strategical shows that the university has projects that show indicators of similarity according to the strategic areas policy, established by the Brazilian Government.

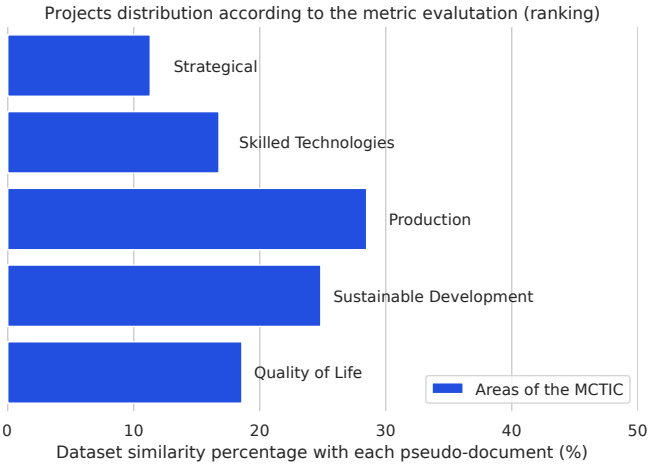

Fig. 6: Dataset ranking.

Looking further, we also investigate the distribution of projects into the most effective areas, according to the adopted methodology. In Figure 7, it is shown the subarea distribution of projects of the area Quality of Life. It can be seen in this graph that the university has a great number of projects associated with the subarea Assistive Technologies, but Sanitation needs more attention. In this case, the university may produce internal policies in order to improve the participation of researchers in this area.

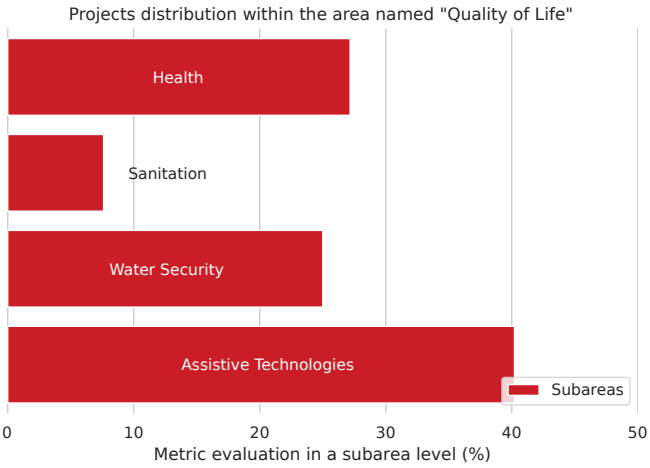

Fig. 7: Subarea level dataset ranking (Quality of Life).

Moreover, when looking at the generated distribution of the great area Skilled Technologies, shown by the Figure 8, the subarea Biotechnology carries the most of the projects, while Nanotechnology has the least number of associated projects.

VI. Final Remarks

This paper proposed a methodology using text indexing techniques, such as Latent Semantic Indexing, to automatically classify documents, composed of title and abstract of research projects from a sample of the Federal University of Tocantins and associated with the innovation policy proposed by the Brazilian Ministry of Science and Technology.

In this way, it gives almost instantly the profile of the current state of the research potentials of the university when presented a public innovation policy. It is required, though, that the policy should be organized into areas and subareas

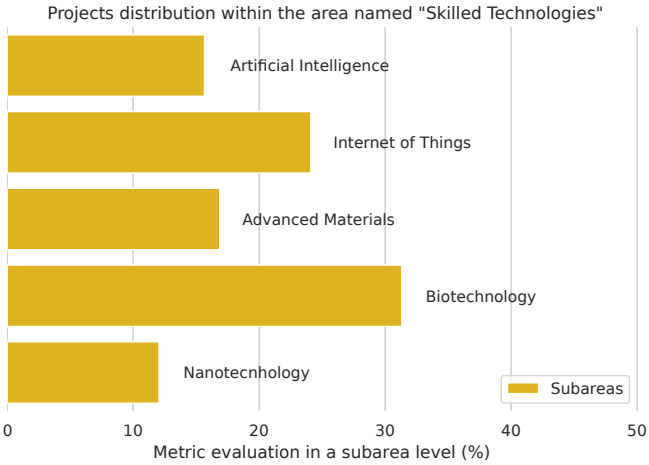

Fig. 8: Subarea level dataset ranking (Skilled Technologies).

with terms that better represent the description of such specific area. In this way, the proposed approach is capable of giving some proximity of projects into the subareas, using the LSI indexing automatically.

The proposed methodology was applied to a sample of projects from five graduate programs of the Federal University. The sample consisted of about 200 projects with title and abstract. When compared with the subareas terms, it could easily be seen that the sample shows a potential of innovation strongly centered at the great areas Production and Sustainable Development. On the other hand, the sample showed fewer projects associated with the great area Strategical. In both cases, the university can produce internal policies to either promote more projects into the strongest areas as well as to build strategies to act more effectively into the less active areas.

The main advantage of the proposed study is that it could be adapted by any research institution that wants to quickly have an overview of actuation based on the described research projects when public policies are presented by government or industry strategies. Unfortunately, this approach depends on what is written in the projects. Therefore, it is still necessary to fine-tune management to accompany the development and impacts of such projects.

As future research, it is intended to include more fields to produce a more effective way to manage the research impacts, such as to relate the projects with patents as well as papers published related to the projects.

REFERENCES

[1] C. Bacchi, *Analysing policy*. Pearson Higher Education AU, 2009, pp. X–XXI.

[2] K. Samset and T. Christensen, "Ex ante project evaluation and the complexity of early decision-making," *Public Organization Review*, vol. 17, no. 1, p. 1–17, 2015.

[3] G. Fahrenkrog, W. Polt, J. Rojo, A. Tübke, K. Zinöcker, S. A. ETH, M. Boden, S. Bührer, R. Cowan, J. Eaton *et al.*, "Rtd evaluation toolbox," *Assessing the Socio-Economic Impact of RT D-P olicies, Seville, European Commission-Joint Research Centre, IPTS*, pp. 25–27, 2002.

[4] K. Flanagan, E. Uyarra, and M. Laranja, "Reconceptualising the 'policy mix' for innovation," *Research Policy*, vol. 40, no. 5, pp. 702–713, 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0048733311000345

[5] MCTIC, "Portaria nº 1.122, de 19 de marÇo de 2020," 2020. [Online]. Available: https://www.in.gov.br/en/web/dou/-/portaria-n-1.122-de-19-de-marco-de-2020-249437397

[6] R. X. Coutinho, E. S. Dávila, W. M. dos Santos, J. B. Rocha, D. O. Souza, V. Folmer, and R. L. Puntel, "Brazilian scientific production in science education," *Scientometrics*, vol. 92, no. 3, pp. 697–710, 2012.

[7] S. Deerwester, S. Dumais, T. Landauer, G. Furnas, and R. Harshman, "Indexing by latent semantic analysis," *J. Am. Soc. Inf. Sci.*, vol. 41, pp. 391–407, 1990.

[8] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008, pp. 26–27, 32–34.

[9] B. Rosario, "Latent semantic indexing: An overview," *Techn. rep. INFOSYS*, vol. 240, pp. 1–16, 2000.

[10] S. Sarica and J. Luo, "Stopwords in technical language processing," *Plos one*, vol. 16, no. 8, p. e0254937, 2021.

[11] A. G. Jivani *et al.*, "A comparative study of stemming algorithms," *Int. J. Comp. Tech. Appl*, vol. 2, no. 6, pp. 1930–1938, 2011.

[12] C. Zhai and S. Massung, *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining*. Association for Computing Machinery and Morgan & Claypool, 2016, pp. 147–153, 90–108.

[13] W. B. Croft, D. Metzler, and T. Strohman, *Search engines: Information retrieval in practice*. Addison-Wesley Reading, 2010, vol. 520, pp. 237–288, 451–452.

[14] R. Baeza-Yates, B. Ribeiro-Neto *et al.*, *Modern information retrieval*. ACM press New York, 1999, vol. 463.

[15] S. C. Deerwester, S. T. Dumais, G. W. Furnas, R. A. Harshman, T. K. Landauer, K. E. Lochbaum, and L. A. Streeter, "Computer information retrieval using latent semantic structure," Jun. 13 1989, uS Patent 4,839,853.

[16] J. L. Bigelow, A. Edwards, and L. Edwards, "Detecting cyberbullying using latent semantic indexing," in *Proceedings of the first international workshop on computational methods for CyberSafety*, 2016, pp. 11–14.

[17] H. Chen, B. Martin, C. M. Daimon, and S. Maudsley, "Effective use of latent semantic indexing and computational linguistics in biological and biomedical applications," *Frontiers in physiology*, vol. 4, p. 8, 2013.

[18] D. Thorleuchter and D. Van den Poel, "Improved multilevel security with latent semantic indexing," *Expert Systems with Applications*, vol. 39, no. 18, pp. 13 462–13 471, 2012.

[19] R. Homayouni, K. Heinrich, L. Wei, and M. W. Berry, "Gene clustering by latent semantic indexing of medline abstracts," *Bioinformatics*, vol. 21, no. 1, pp. 104–115, 2005.

[20] Q. Wang, J. Xu, H. Li, and N. Craswell, "Regularized latent semantic indexing: A new approach to large-scale topic modeling," *ACM Transactions on Information Systems (TOIS)*, vol. 31, no. 1, pp. 1–44, 2013.

[21] H. Elghazel, A. Aussem, O. Gharroudi, and W. Saadaoui, "Ensemble multi-label text categorization based on rotation forest and latent semantic indexing," *Expert Systems with Applications*, vol. 57, pp. 1–11, 2016.

[22] R. V. R. D. Souza, "Análise textual dos projetos de inovação da fundação universidade federal do tocantins à luz das áreas e setores tecnológicos prioritários do ministério da ciência, tecnologia, inovação e comunicação." Master's thesis, Universidade Federal do Tocantins, Tocantins, Brazil, 2021.

[23] MCTIC, "Estratégia nacional de ciência, tecnologia e inovação: 2016-2022," 2016. [Online]. Available: http://www.finep.gov.br/images/a-finep/Politica/16_03_2018_Estrategia_Nacional_de_Ciencia_Tecnologia_e_Inovacao_2016_2022.pdf